

AI 實作課程

Azure ML X Python

Kristen Chan

2019.05.18

Agenda

- Kaggle
- Microsoft Azure Machine Learning
- Python



* kaggle

Hello !

I am Kristen Chan

Data Scientist

E-Commerce / Telecom

R-Ladies Taipei

Co-Organizer [link](#)



Kaggle Introduction

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 'ggle' part.

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 rectangle), 'Datasets', 'Kernels', 'Discussion', 'Learn', and more. Below the navigation bar is a 'Newsfeed' section titled 'Welcome Kristen Chan'. It includes a brief introduction about the newsfeed and a post from user 'Yakin' about a dataset. To the right of the newsfeed is a profile card for 'Kristen Chan' and a 'Featured Job' section for 'Invisibly' hiring a Data Scientist in Palo Alto, CA, USA. At the bottom of the page, there is a 'Your Competitions' section.

kaggle Search Competitions Datasets Kernels Discussion Learn ...

Newsfeed

Welcome Kristen Chan

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.

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

Kristen Chan Joined 2 years ago

Novice

Featured Job

Invisibly is hiring Data Scientist
Palo Alto, CA, USA

Your Competitions

Deepak Malpani · Follow

KristenChan 7

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 '經典' (Classic). An arrow points from a pink box to this highlighted entry. Below it, another arrow points to a blue box containing the text '目前有 18 個正在進行中的比賽' (Currently there are 18 active competitions).

Competitions

General InClass Documentation 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 8

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-instruction 'Start here! Predict survival on the Titanic and get familiar with ML basics' follows. On the left, there's a 'Kaggle' logo and the number '10,049 teams · Ongoing'. Below this, a navigation bar includes 'Overview' (which is underlined), 'Data', 'Kernels', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and a prominent blue 'Submit Predictions' button. The main content area starts with a section titled 'Description' containing the text 'Start here if...'. It then moves to 'Evaluation', which states: 'You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.' Next is 'Tutorials', followed by 'Frequently Asked Questions'. The 'Competition Description' section details the sinking of the RMS Titanic on April 15, 1912, where it collided with an iceberg, killing 1502 out of 2224 passengers and crew. It notes that safety regulations for ships were improved after the tragedy. The final part of the description explains the goal of the challenge: to analyze passenger data to predict survival based on various factors like age, gender, and class.

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 |

891 X 12

418 X 11

418 X 2

| | | |
|----------------|----------|--|
| 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

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/>

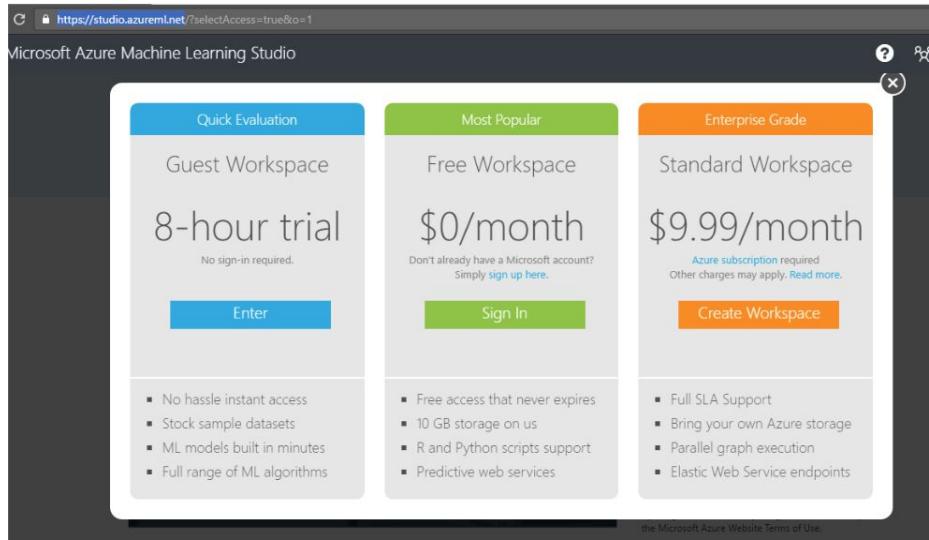
The screenshot shows the main landing page of the Azure Machine Learning Studio. At the top, there's a navigation bar with three horizontal bars, the text "Microsoft Azure Machine Learning Studio", and icons for help, user profile, and sign in. Below the header is a large dark banner featuring a blue 3D tetrahedron icon and the text "Azure Machine Learning service". A yellow "New!" badge is in the top right corner of the banner. Below the banner, the text "Try it today!" is displayed. 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 items like "Azure Machine Learning Studio R Runtime Upgrade" and "Mining Campaign Funds".



Krisjen Chian 22

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 25

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 26

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

DOWNLOAD DELETE OPEN IN NOTEBOOK GENERATE DATA ACCESS ADD TO PROJECT

NEW



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 28

建立 Experiment

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

experiments

PROJECTS EXPERIMENTS NEW

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:

Kristen Chian 29

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 points to the title "Kaggle Titanic Easy" at the top of the workspace.

[Canvas]
編輯實驗
並放上你需要的 Module

To create your experiment, drag items from the sidebar into the canvas. You can also use the "Drag Items Here" placeholder.

The workspace itself is a "Mini Map" view showing a flowchart of data processing steps connected by dashed lines. A "Quick Help" button is located in the bottom right corner of the workspace area.

Properties Project

Experiment Properties

Status Code InDraft

Summary

Description

Quick Help

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

Kristen Chian 30

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 DISCARD CHANGES SET UP WEB SERVICE PUBLISH TO GALLERY

The screenshot shows the Microsoft 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, DISCARD CHANGES, SET UP WEB SERVICE, and PUBLISH TO GALLERY. The top right corner shows the user 'Kristen-Free-Workspace' with profile icons for help, users, smiley face, and a person. The status bar at the bottom indicates 'In draft' and shows 'Properties' and 'Project' tabs. On the far right, there's a 'Quick Help' section and a large blue watermark for 'Kristen Chian 31'.

介面介紹

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

Search experiment items

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

Deprecated

Drag Items Here

In draft

[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

Mini Map

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

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 32

The screenshot illustrates the Microsoft Azure Machine Learning Studio interface. On the left, a sidebar lists experiment items like Saved Datasets, Data Format Conversions, and Machine Learning modules. The main workspace shows a 'Kaggle Titanic Easy' experiment with a 'Mini Map' tool. A callout box highlights the 'Properties' panel on the right, which contains sections for Experiment Properties (Status Code: InDraft), Summary (a placeholder for a brief description), and Description (a placeholder for a detailed description). A note in the center states: '當點選 Canvas 上的 Module 時, Properties 會出現對應的參數供調整' (When you select a module on the Canvas, the Properties will appear to provide corresponding parameters for adjustment). Another note below it says: '[Note] Quick Help 可以幫助你更深入了解個參數意義' ([Note] Quick Help can help you understand the meaning of each parameter more deeply). The bottom navigation bar includes links for RUN HISTORY, SAVE, DISCARD CHANGES, SET UP WEB SERVICE, and PUBLISH TO GALLERY.

介面介紹

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

[控制執行] 儲存、執行、發布實驗...

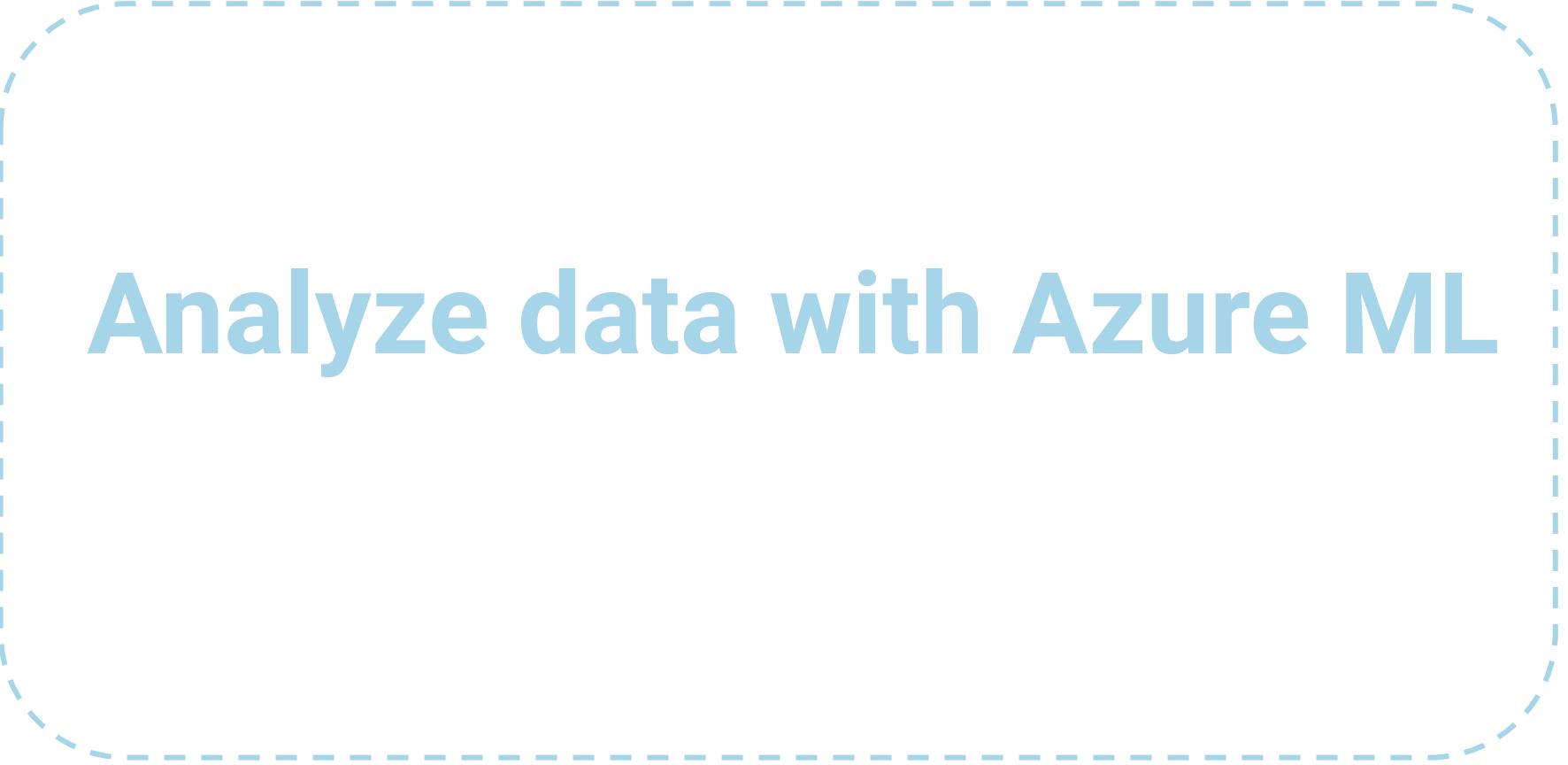
Drag Items Here

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

Mini Map

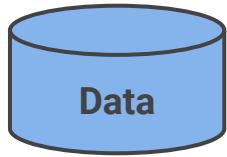
D NEW RUN HISTORY SAVE DISCARD CHANGES SET UP WEB SERVICE PUBLISH TO GALLERY

Kristen Chian 33



Analyze data with Azure ML

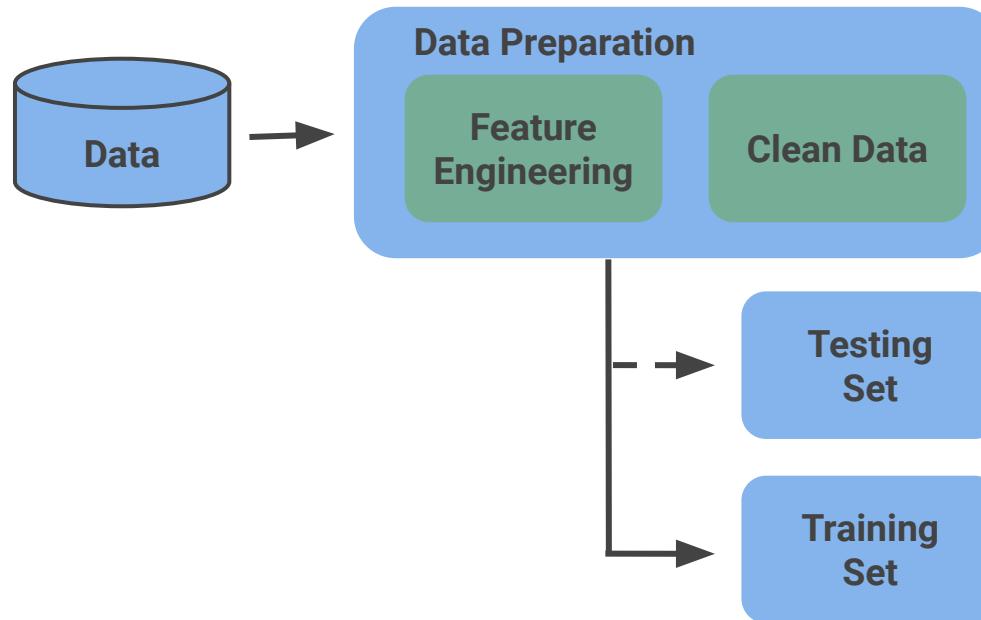
Data processing pipeline



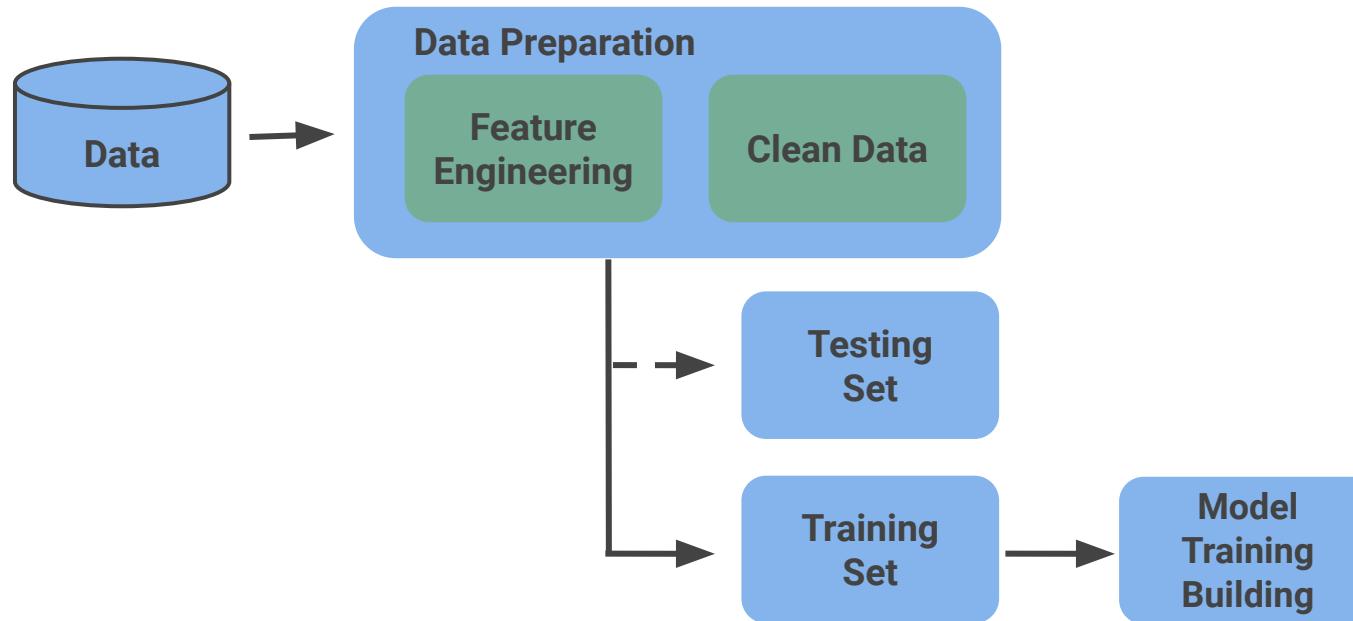
Data processing pipeline



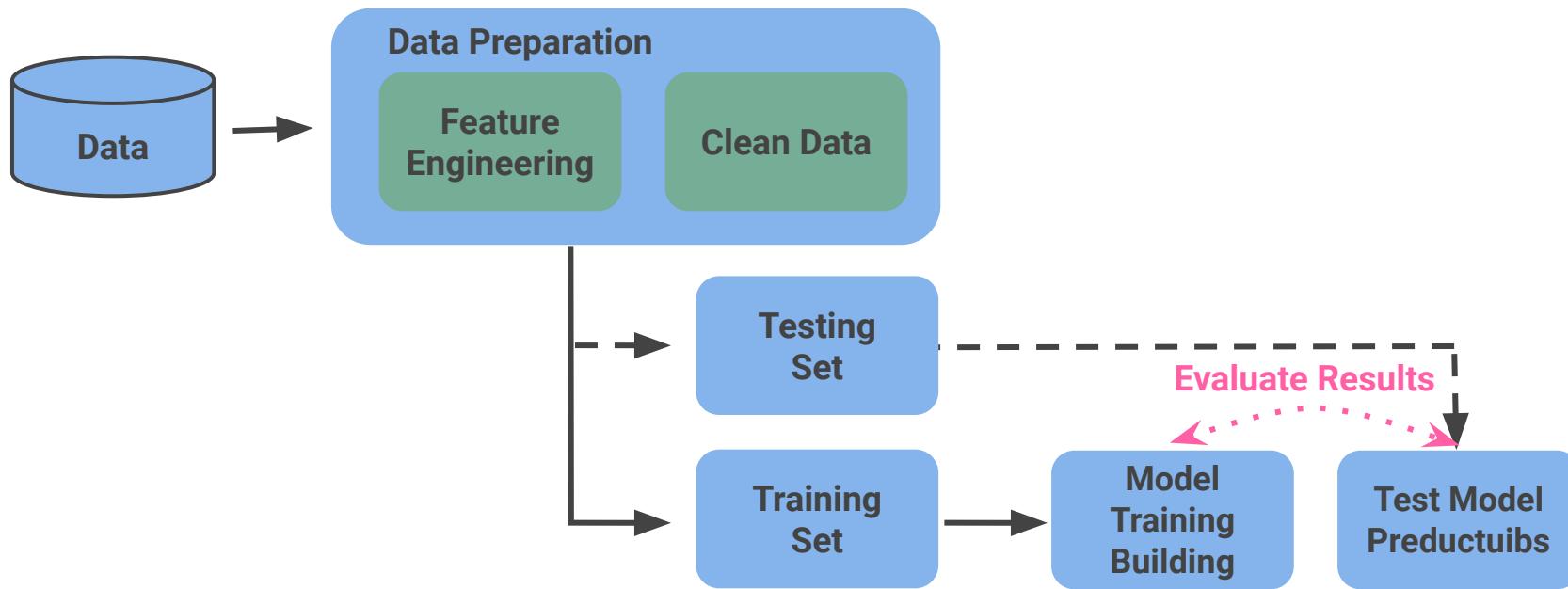
Data processing pipeline



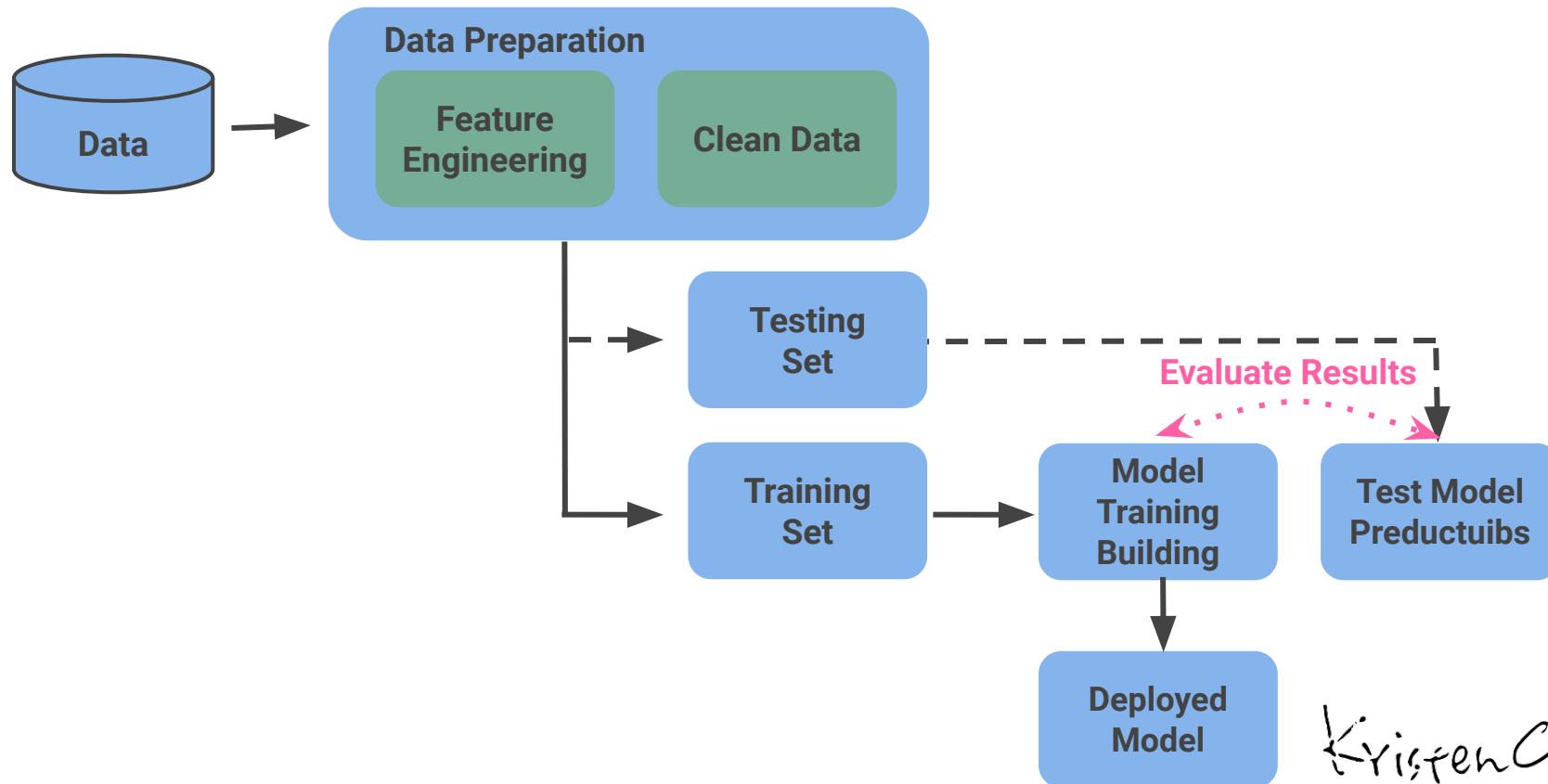
Data processing pipeline



Data processing pipeline



Data processing pipeline



All Data Sets

剛剛Kaggle 的資料集



| 資料集 | 資料維度 | 剛剛Kaggle 的資料集 | | |
|--------------------------|----------|----------------|-------|---------------------------|
| 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 |

Yi Chen Chiu 41

[Method 1] Access Data

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

Search experiment items

Saved Datasets

My Datasets

- Titanic_Test.csv
- Titanic_Train.csv**

Drag Items Here

匯入資料
→ 從 Module 選擇, Saved Datasets 中的 [Titanic_Train.csv]

Summary

Description

Quick Help

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

Kristen Chaw 42

[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 Machine Learning Studio interface. On the left, there's a sidebar with various icons and a list of datasets and operations. In the center, a 'Mini Map' shows the 'Titanic_Train.csv' dataset. A context menu is open over the dataset icon, with the 'Visualize' option highlighted. The 'Properties' panel on the right displays information about the dataset, including its submission details and file size.

In draft

Kristen-Free-Workspace

Properties Project

Titanic_Train.csv

SUBMITTED BY sinue625

SIZE 59.8 KB

FORMAT GenericCSV

CREATED ON 2/12/2019 ...

View dataset

Quick Help

Kaggle Titanic Train

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

PUBLISH TO GALLERY

SET UP WEB SERVICE

Kristen Chiau 43

[Method 2] Access Data

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace ▾ ? 🌐 😊 🚙

Project Import Data Wizard

Azure Blob Storage

Authentication type: Storage Account

Account name: (red exclamation mark)

Account key: (red exclamation mark)

Path to container, directory: (red exclamation mark)

Blob file format: CSV

File has header row

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.
[\(more help...\)](#)

匯入資料

→ 從 Module 選擇, Data Input and Output 中的 [Import Data]

Import Data

Enter Data Manually

Export Data

Import Data

Load Trained Model

Unpack Zipped Datasets

Data Transformation

Feature Selection

Machine Learning

OpenCV Library Modules

Python Language Modules

R Language Modules

Statistical Functions

Text Analytics

Mini Map

Import Data

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

Kristen Chian 44

[Method 2] Access Data

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The left sidebar contains navigation icons and a search bar. The main area is titled "Kaggle Titanic Easy" and shows a "Import Data" step with a red exclamation mark. The "Properties" tab is selected, displaying configuration options for the import source (Azure Blob Storage), authentication type (Storage Account), account name, account key, and path to container. A blue arrow points from the "Import Data" step in the canvas to the "Import Data" section in the properties panel.

Kaggle Titanic Easy

In draft

Properties Project

Import Data

Launch Import Data Wizard

Data source: Azure Blob Storage

Authentication type: Storage Account

Account name: [redacted]

Account key: [redacted]

Path to container, directory...: [redacted]

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

Statistical Functions

Text Analytics

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

Kristen Chiau 45

https://raw.githubusercontent.com/kristenchan/Sharing/master/Kaggle_Titanic/Titanic_Train.csv

[Method 2] Access Data

The screenshot illustrates the "Import Data" wizard in Azure Machine Learning Studio. It consists of three main panels:

- Left Panel (Main Studio View):** Shows the Azure Machine Learning Studio interface with various modules like Feature Selection, Machine Learning, and Text Analytics. A pink callout box with the text "資料有 Header 就要打勾" (Checkmark if the data has a header) points to the "CSV or TSV has header row" checkbox in the "Import Data" dialog.
- Middle Panel (Import Data Wizard Step 1):** Titled "Choose data source". It lists several options: "Web URL via HTTP" (selected and highlighted with a pink box), Hive Query, Azure SQL Database, Azure Table, Azure Blob Storage, Data Feed Provider, On-Premises SQL Database (Preview Feature), and Azure DocumentDB. A pink circle labeled "1" is on the "Launch Import Data Wizard" button.
- Right Panel (Import Data Wizard Step 2):** Titled "Import Data". It shows the configuration for "Web URL via HTTP": "Data source URL" is set to "https://raw.githubusercontent.com/kristenchan/Sharing/master/Kaggle_Titanic/Titanic_Train.csv", "Data format" is "CSV", and "CSV or TSV has header row" is checked. A pink circle labeled "2" is on the "Data source URL" field. A pink arrow points from the "Data source URL" field in the right panel to the "Data source URL" field in the middle panel.

A dashed blue arrow at the top indicates the flow from the main studio view to the import wizard.

Kristen Chan 46

[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 Machine Learning studio interface. On the left, there's a sidebar with various icons and a list of modules: Saved Datasets, Data Format Conversions, Data Input and Output (selected), Enter Data Manually, Export Data, Import Data, Load Trained Model, Unpack Zipped Datasets, Data Transformation, Feature Selection, Machine Learning, OpenCV Library Modules, Python Language Modules, R Language Modules, Statistical Functions, and Text Analytics. In the center, a workflow canvas displays a single step labeled "Import Data". Below the canvas is a "Mini Map" showing the same step. At the bottom of the screen, there's a toolbar with buttons for NEW, RUN HISTORY, SAVE, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY. A pink circle with the number "1" highlights the "Run" button. On the right side, there are sections for Properties, Project, Import Data (with fields for Data source, URL, and format), and Quick Help (describing the Import Data step). The top right corner shows the workspace name "Kristen-Free-Workspace" and user profile information.

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...)

Run selected

Kristen Chian 47

[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 Machine Learning 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 experiment canvas is the 'Properties' pane, which displays the status 'Finished running' and the 'Import Data' configuration. The 'Data source URL' is set to 'Web URL via HTTP' with the value 'https://raw.githubusercontent.com'. The 'Data format' is set to 'CSV'. The 'START TIME' is '9/3/2018 1...' and the 'END TIME' is '9/3/2018 1...'. The 'ELAPSED TIME' is '0:00:09.990' and the 'STATUS CODE' is 'Finished'. The 'STATUS DETAILS' is 'None'. At the bottom of the Properties pane, there is a 'Quick Help' section with a brief description of the Import Data step.

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)

KristenChian 48

Access Data

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

Kaggle Titanic Easy > Import Data > Results dataset

rows 891

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 |
|-------|----------|----------------|--------|-----|---------------|-------------|-----------|--------|
| Enter | 0 | 3 | male | 22 | 1 | 0 | 7.25 | S |
| Explo | 1 | 1 | female | 38 | 1 | 0 | 71.2833 | C |
| Impor | 1 | 3 | female | 26 | 0 | 0 | 7.925 | S |
| Load | 1 | 1 | female | 35 | 1 | 0 | 53.1 | S |
| Unpa | 0 | 3 | male | 35 | 0 | 0 | 8.05 | S |
| Ente | 0 | 3 | male | 0 | 0 | 0 | 8.4583 | Q |
| Expl | 0 | 1 | male | 54 | 0 | 0 | 51.8625 | S |
| Impo | 0 | 3 | male | 2 | 3 | 1 | 21.075 | S |
| Loa | 1 | 3 | female | 27 | 0 | 2 | 11.1333 | S |
| Unpa | 1 | 2 | female | 14 | 1 | 0 | 30.0708 | C |
| Ent | 1 | 3 | female | 4 | 1 | 1 | 16.7 | S |
| Expl | 1 | 1 | female | 58 | 0 | 0 | 26.55 | S |
| Ente | 0 | 2 | male | 20 | 0 | 0 | 0.25 | S |

To view, select a column in the table.

Statistics

Visualizations

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

Kristen Chau 49

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's a sidebar with icons for various operations like Data Transformation, Filter, Learning with Counts, and Manipulation. Under Manipulation, the 'Select Columns in Dataset' module is highlighted with a red box and a pink arrow pointing from the main text above. In the center, a flowchart shows a sequence of modules: Import Data, followed by Select Columns in Dataset (which has a circled '1' below it), and then another Select Columns in Dataset module (also with a circled '1'). A 'Mini Map' window on the left shows the same sequence. On the right, the properties pane is open for the 'Select Columns in Dataset' module, showing its configuration options. The status bar at the bottom includes buttons for NEW, RUN HISTORY, SAVE, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY.

Kristen-Free-Workspace

In draft

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

Import Data

Select Columns in Dataset

Mini Map

Import Data

Select Columns in Dataset

1

1

NEW

RUN HISTORY

SAVE

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

Quick Help

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

Kristen Chian 50

篩選資料

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

In draft

Draft saved at 下午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...)

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

KristenChian 51

```
graph TD; ImportData[Import Data] --> SelectColumns[Select Columns in Dataset];
```

篩選資料

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

In draft Draft saved at 午後1:40:54

Properties Project

Select columns

Survived, Pclass, Sex, Age

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

2 Launch column selector

要的選過來

3 選完記得打勾

PassengerId Name Sex SibSp Parch Ticket Fare Cabin Embarked

All Types search columns

All Types search columns

9 columns available 8 columns selected

Quick Help

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

(more help...)

Kristen-Free-Workspace

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

確認資料 – 確定資料型態

Kaggle Titanic Input3

Kaggle Titanic Input3 > Select Columns in Dataset > Results dataset

rows: 891 columns: 4

| | Survived | Pclass | Sex | Age |
|---|----------|--------|--------|-----|
| 0 | 0 | 3 | male | 22 |
| 1 | 1 | 1 | female | 38 |
| 1 | 1 | 3 | female | 26 |
| 1 | 1 | 1 | female | 35 |
| 0 | 0 | 3 | male | 35 |
| 0 | 0 | 3 | male | |
| 0 | 0 | 1 | male | 54 |
| 0 | 0 | 3 | male | 2 |
| 1 | 1 | 3 | female | 27 |
| 1 | 1 | 2 | female | 14 |
| 1 | 1 | 3 | female | 4 |
| 1 | 1 | 1 | female | 58 |
| 0 | 0 | 3 | male | 20 |
| 0 | 0 | 3 | male | 39 |
| 0 | 0 | 3 | female | 14 |
| 1 | 1 | 2 | female | 55 |
| 0 | 0 | 3 | male | 2 |
| 1 | 1 | 2 | male | |
| 0 | 0 | 3 | female | 31 |
| 1 | 1 | 3 | female | |
| 0 | 0 | 2 | male | 35 |
| 1 | 1 | 2 | male | 34 |
| 1 | 1 | 3 | female | 15 |
| 1 | 1 | 3 | male | 20 |

Note Pclass 的資料只有 1, 2, 3 三種
資料類型應該是類別型

Statistics

| Feature Type | Numeric Feature |
|--------------------|-----------------|
| Mean | 2.3086 |
| Median | 3 |
| Min | 1 |
| Max | 3 |
| Standard Deviation | 0.8361 |
| Unique Values | 3 |
| Missing Values | 0 |

Visualizations

Pclass Histogram

frequency

Pclass

Yisten

Note Pclass 的資料只有 1, 2, 3 三種
資料類型應該是類別型

確認資料 - 確定資料型態

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3

Import Data → Select Columns in Dataset → Edit Metadata

Selected columns: Pclass, Sex

Data Type : String

1. Launch column selector

2. Available columns: Survived, Age

3. Selected columns: Pclass, Sex

BY NAME WITH RULES

AVAILABLE COLUMNS: All Types | search columns

SELECTED COLUMNS: All Types | search columns

Pclass, Sex

2 columns available

2 columns selected

Mini Map

Import Data → Select Columns in Dataset → Edit Metadata

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

Kristen-Free-Workspace Properties Project

Column

Selected columns: Pclass, Sex

Launch column selector

Data type: String

Categorical: Unchanged

Fields: Unchanged

New column names: []

START TIME: 5/17/2019 ...
END TIME: 5/17/2019 ...
ELAPSED TIME: 0.0002.047
STATUS CODE: Finished
STATUS DETA...: None

View output log

Quick Help

Edits metadata associated with columns in a dataset. Formerly known as Metadata Editor.

(more help...)

Kristen Chian 54

確認資料 - 確定資料型態

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3 > Edit Metadata > Results dataset

rows: 891 columns: 4

view as: **Pclass**

| | Survived | Pclass | Sex | Age |
|----|----------|--------|--------|-----|
| 0 | 0 | 3 | male | 22 |
| 1 | 1 | 1 | female | 38 |
| 2 | 1 | 3 | female | 26 |
| 3 | 1 | 1 | female | 35 |
| 4 | 0 | 3 | male | 35 |
| 5 | 0 | 3 | male | |
| 6 | 0 | 1 | male | 54 |
| 7 | 0 | 3 | male | 2 |
| 8 | 1 | 3 | female | 27 |
| 9 | 1 | 2 | female | 14 |
| 10 | 1 | 3 | female | 4 |
| 11 | 1 | 1 | female | 58 |
| 12 | 0 | 3 | male | 20 |
| 13 | 0 | 3 | male | 39 |
| 14 | 0 | 3 | female | 14 |
| 15 | 1 | 2 | female | 55 |
| 16 | 0 | 3 | male | 2 |
| 17 | 1 | 2 | male | |
| 18 | 0 | 3 | female | 31 |
| 19 | 1 | 3 | female | |
| 20 | 0 | 2 | male | 35 |
| 21 | 1 | 2 | male | 34 |
| 22 | 1 | 3 | female | 15 |
| 23 | 1 | 1 | male | 29 |

Note Pclass 被改為類別類型

Statistics

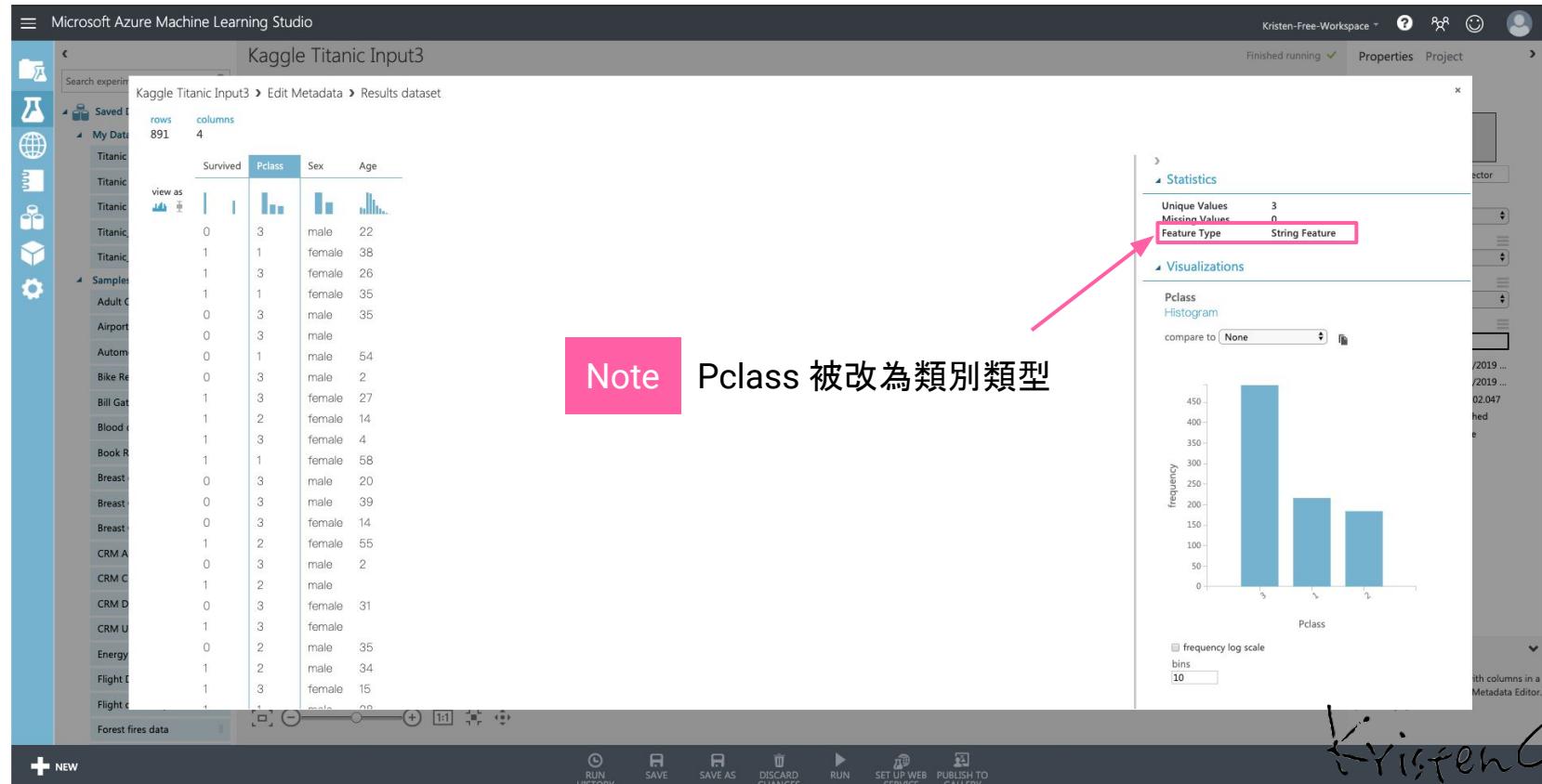
Unique Values: 3
Missing Values: 0
Feature Type: String Feature

Visualizations

Pclass Histogram

frequency

bins: 10



KristenChen 55

清理資料

Microsoft Azure Machine Learning Studio

清理欄位

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

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there is a navigation pane with sections like 'Saved Datasets', 'Transforms' (selected), 'Data Transformation', 'Manipulation', and 'Text Analytics'. Under 'Manipulation', 'Clean Missing Data' is selected. The main workspace displays a data flow diagram. It starts with 'Select Columns in Dataset', followed by 'Edit Metadata', and then 'Clean Missing Data'. The 'Clean Missing Data' module has two output ports, labeled 1 and 2. A 'Mini Map' window at the bottom left shows the same flow. On the right side, there is a configuration panel for the 'Clean Missing Data' module, which includes settings for 'Maximum missing value ratio' (set to 1), 'Cleaning mode' (set to 'Custom substitution value'), 'Replacement value' (set to 0), and a checkbox for 'Generate missing val...'. A 'Quick Help' tooltip is also visible. At the bottom, there are standard studio navigation buttons: RUN HISTORY, SAVE, SAVE AS, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY. A watermark 'KrisTenChen 56' is in the bottom right corner.

Import Data

Select Columns in Dataset

Edit Metadata

Clean Missing Data

1 2

Mini Map

Quick Help

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

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

KrisTenChen 56

[Method 1] 清理資料 -- 整筆刪掉

Kaggle Titanic Input3

In draft

Draft saved at: 午後12:09:44

Properties Project

Selected columns:
All columns

Replace using MICE
Custom substitution value
Replace with mean
Replace with median
Replace with mode
Remove entire row (checked)
Remove entire column
Replace using Probabilistic PCA

Replace using MICE
Custom substitution value
Replace with mean
Replace with median
Replace with mode
Remove entire row (checked)
Remove entire column
Replace using Probabilistic PCA

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

Mini Map

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

Kristen-Chian 57

[Method 1] 清理資料 -- 整筆刪掉

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3

Search ex: Kaggle Titanic Input3 > Clean Missing Data > Cleaned dataset

ROWS 714

Note 剩下有 714 筆資料

| | Survived | Pclass | Sex | Age |
|---|----------|--------|--------|-----|
| T | 0 | 3 | male | 22 |
| T | 1 | 1 | female | 38 |
| T | 1 | 3 | female | 26 |
| A | 1 | 1 | female | 35 |
| A | 0 | 3 | male | 35 |
| A | 0 | 1 | male | 54 |
| B | 0 | 3 | male | 2 |
| B | 1 | 3 | female | 27 |
| B | 1 | 2 | female | 14 |
| B | 1 | 3 | female | 4 |
| B | 1 | 1 | female | 58 |
| B | 0 | 3 | male | 20 |
| B | 0 | 3 | male | 39 |
| B | 0 | 3 | female | 14 |

To view, select a column in the table.

CRM Appetency Labels ...

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

KristenChian 58

[Method 2] 清理資料 -- 數值型資料用中位數取代

處理數值型資料

1 Launch column selector

2 Survived Age

KristenChian 59

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a navigation sidebar with various options like 'Saved Datasets', 'Transforms', 'Data Transformation', and 'Text Analytics'. In the center, a workflow diagram is displayed with steps: 'Import Data' followed by 'Select Columns in Dataset'. A modal window titled 'Select columns' is open, showing two lists: 'AVAILABLE COLUMNS' (listing 'Pclass', 'Sex') and 'SELECTED COLUMNS' (listing 'Survived', 'Age'). A pink box labeled '2' highlights the 'Survived' and 'Age' entries in the 'SELECTED COLUMNS' list. A pink arrow points from this box to the 'Launch column selector' button in the top right corner of the modal. The top right also shows settings for 'Clean Missing Data', including 'Selected columns' set to 'Survived,Pclass,Sex,Age', 'Minimum missing value r...' set to '0', and 'Maximum missing value r...' set to '1'. The bottom right of the screen has a watermark reading 'KristenChian 59'.

[Method 2] 清理資料 -- 數值型資料用中位數取代

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, the navigation pane includes sections for Saved Datasets, Samples, Transforms, Data Transformation, and Text Analytics. A search bar at the top right contains the text 'cle'. The main workspace displays a data flow diagram:

```
graph TD; ImportData[Import Data] --> SelectColumns[Select Columns in Dataset]; SelectColumns --> EditMetadata[Edit Metadata]; EditMetadata --> CleanMissingData[Clean Missing Data];
```

The 'Clean Missing Data' step is highlighted with a red box and has two numbered callouts: ① and ②. A pink arrow points from the explanatory text below to the 'Replace with median' option in the context menu.

Properties panel (right side):

- Clean Missing Data**
- Selected columns:** Column names: Survived,Age
- Launch column selector**
- Minimum missing value r...**: 0
- Maximum missing value r...**: 1
- Replace using MICE**
- Custom substitution value**
- Replace with mean**
- Replace with median** (highlighted with a blue border)
- Replace with mode**
- Remove entire row**
- Remove entire column**
- Replace using Probabilistic PCA**

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

Kristen Chan 60

[Method 2] 清理資料 -- 數值型資料用中位數取代

Microsoft Azure Machine Learning Studio

Kaggle Titanic Easy

Finished running ✓ Properties Project

Kaggle Titanic Easy > Clean Missing Data > Cleaned dataset

ROWS 891

Note 維持 891 筆資料

檢查資料處理狀況

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

| A | 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 |
| R | 0 | 0 | 60 |

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 61

[Method 3] 清理資料 -- 類別型資料用眾數取代

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

Kaggle Titanic Input3 In draft Draft saved at 上午12:40:25

Transforms Expe [Clean Missing Data] Expe [Clean Missing Data] 1

Data Transformation Manipulation Clean Missing Data

Text Analytics Preprocess Text

Select columns

AVAILABLE COLUMNS

BY NAME WITH RULES

All Types search columns

Survived
Age

2 columns available

2 columns selected

CREATE

SELECTED COLUMNS

Sex
Pclass

處理類別型資料

1 Launch column selector

2

Columns to be cleaned

Selected columns

Minimum missing value ratio: 0

Maximum missing value ratio: 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...)

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

KristenChian 62

[Method 3] 清理資料 -- 類別型資料用眾數取代

Kaggle Titanic Input3

```
graph TD; Import[Import Data] --> Select[Select Columns in Dataset]; Select --> Edit[Edit Metadata]; Edit --> Clean1[Clean Missing Data]; Clean1 --> Clean2[Clean Missing Data];
```

In draft

Saving...

Properties Project

Clean Missing Data

Selected columns: Column names: Pclass,Sex

Launch column selector

Minimum missing value r... 0

Replace using MICE
Custom substitution value
Replace with mean
Replace with median
Replace with mode
Remove entire row
Remove entire column
Replace using Probabilistic PCA
Generate missing val...

處理 Missing 方法
在 Cleaning mode 中選 [Replace with mode]：當類別欄位出現遺失值時，用眾數取代整筆

處理 Missing 方法

在 Cleaning mode 中選
[Replace with mode] : 當類別欄位出現遺失值時,
用眾數取代整筆

分割資料 – Train Set / Test Set

Microsoft Azure Machine Learning Studio

分割資料

→ 從 Module 選擇, Data Transformation 中 Sample and Split 的 [Split Data]

The screenshot shows a data flow in the Microsoft Azure Machine Learning Studio. The process starts with 'Import Data' (Titanic Dataset), followed by 'Select Columns in Dataset', 'Edit Metadata', and 'Clean Missing Data'. The output from 'Clean Missing Data' splits into two parallel paths, each passing through another 'Clean Missing Data' module. Finally, both paths converge at a 'Split Data' module, which is highlighted with a blue border. The 'Split Data' module has two outputs, labeled 1 and 2. A 'Mini Map' window in the bottom-left corner provides a overview of the entire flow. On the right side, there is a configuration panel for the 'Split Data' module, showing settings for 'Splitting mode' (Split Rows), 'Fraction of rows in the first set' (0.7), and 'Randomized split' (checked). A 'Quick Help' section at the bottom right explains the purpose of the module.

Import Data

Select Columns in Dataset

Edit Metadata

Clean Missing Data

Clean Missing Data

Split Data

Mini Map

Quick Help

Split the rows of a dataset into two distinct sets
(more help...)

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

Kristen Chan 64

分割資料 – Train Set / Test Set

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3

In draft

Draft saved at 下午2:44:11

Properties Project

Note

70 % ← Train Set
30 % ← Test Set

Import Data → Select Columns in Dataset → Edit Metadata → Clean Missing Data → Clean Missing Data → Split Data

1 624 筆 2 267 筆

Mini Map

Quick Help

Split the rows of a dataset into two distinct sets
(more help...)

Kristen-Free-Workspace

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

建立 Train Model

建立模型

→ 從 Module 選擇, Machine Learning 中 Train 的 [Train Model]

The screenshot shows the Microsoft Azure Machine Learning studio interface. On the left, there is a navigation pane with various categories like Train, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, and more. The 'Train' section is expanded, showing options like Sweep Clustering, Train Anomaly Detection, Train Clustering Model, Train Matchbox Recom..., Train Model, and Tune Model Hyperpara... A 'Train Model' module is selected and highlighted with a red border. The main workspace displays a workflow diagram starting with 'Import Data', followed by 'Select Columns in Dataset', 'Edit Metadata', 'Clean Missing Data', 'Clean Missing Data' (with a feedback loop), and 'Split Data'. The 'Train Model' module is at the end of the flow, receiving input from the 'Split Data' module. A 'Mini Map' window in the bottom-left corner provides a overview of the entire workflow. The top right of the screen has sections for 'Summary' and 'Description' with placeholder text. The bottom right corner features a signature 'Kristen Chan' and the number '66'.

Microsoft Azure Machine Learn...

Train

Text - Trained N-grams mo...

Data Input and Output

Load Trained Model

Data Transformation

Filter

Threshold Filter

Scale and Reduce

Normalize Data

Feature Selection

Permutation Feature Import...

Machine Learning

Score

Assign Data to Clusters

Score Model

Train

Sweep Clustering

Train Anomaly Detectio...

Train Clustering Model

Train Matchbox Recom...

Train Model

Tune Model Hyperpara...

OpenCV Library Modules

Pre-trained Cascade Image...

Text Analytics

Train Vowpal Wabbit Versi...

Train Vowpal Wabbit Versi...

Deprecated

Import Data

Select Columns in Dataset

Edit Metadata

Clean Missing Data

Clean Missing Data

Split Data

Train Model

Mini Map

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

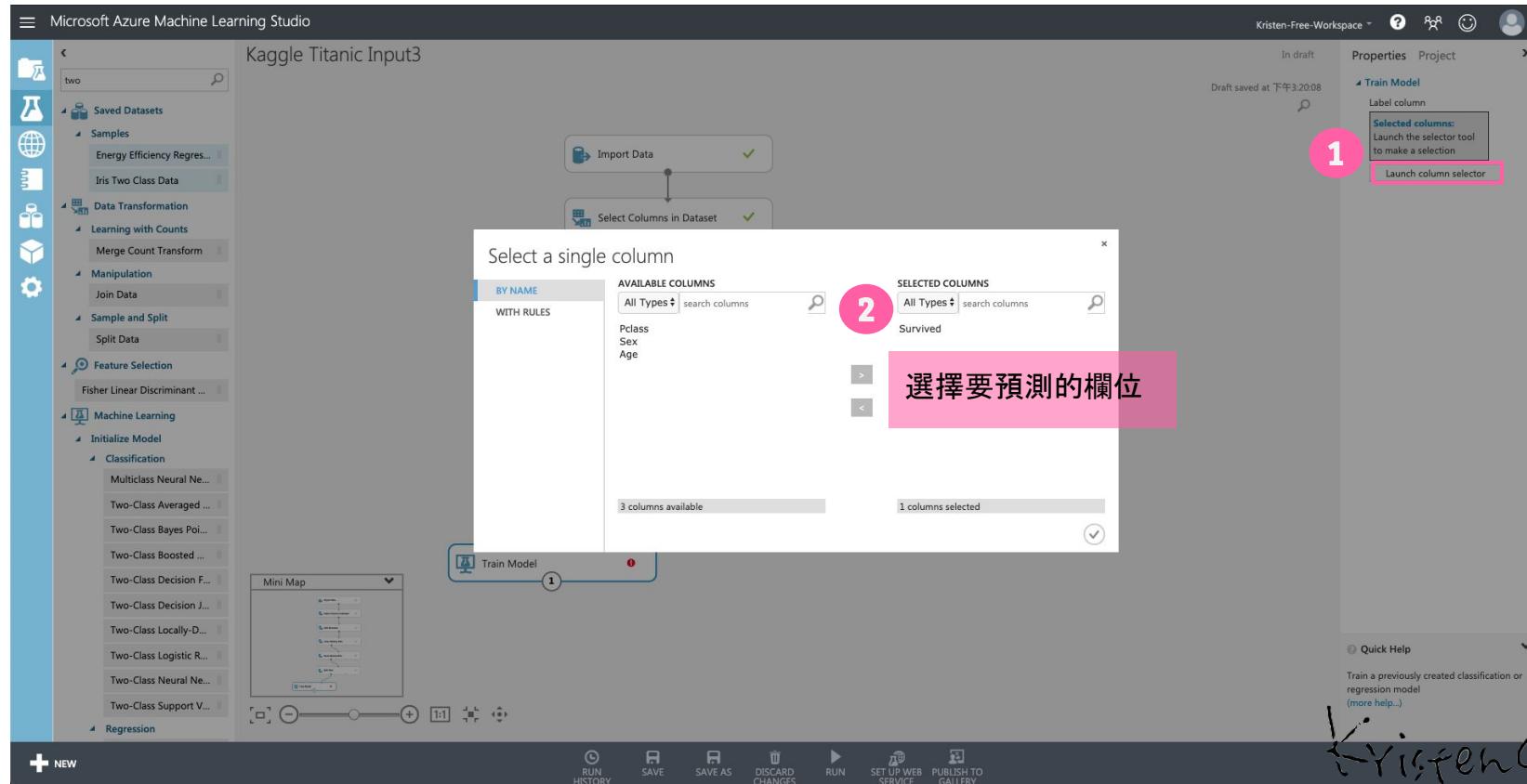
RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

Kristen Chan 66

建立 Train Model



(more help...)

建立 Train Model - 選擇一個分類方法

Microsoft Azure Machine Learning Studio

建立模型

→ 從 Module 選擇, Machine Learning 中 Initialize Model [Classification]

```
graph TD; Import[Import Data] --> Select[Select Columns in Dataset]; Select --> Edit[Edit Metadata]; Edit --> Clean1[Clean Missing Data]; Clean1 --> Clean2[Clean Missing Data]; Clean2 --> Split[Split Data]; Split --> Logistic[Two-Class Logistic Regression]; Logistic --> Train[Train Model];
```

STATUS DATA... None

Prior Run

Summary

Description

Quick Help

Kristen Chan 68

將 Model 套用在 Test Set 中

評分模型

→ 從 Module 選擇, Machine Learning 中 Score 的 [Score Model]

```
graph TD; Import[Import Data] --> Select[Select Columns in Dataset]; Select --> Edit[Edit Metadata]; Edit --> Clean1[Clean Missing Data]; Clean1 --> Clean2[Clean Missing Data]; Clean2 --> Split[Split Data]; Split --> Train[Train Model]; Train --> Score[Score Model];
```

STATUS DATA... None

Prior Run

Summary

Description

Quick Help

Kristen Chau 69

將 Model 套用在 Test Set 中

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3

rows: 267 columns: 6

| | Survived | Pclass | Sex | Age | Scored Labels | Scored Probabilities |
|----------|----------|--------|------|-----|---------------|----------------------|
| 0 | 3 | male | 33 | 0 | 0.093533 | |
| 1 | 1 | female | 38 | 1 | 0.88699 | |
| 1 | 1 | male | 52 | 0 | 0.372824 | |
| 1 | 3 | male | 32 | 0 | 0.094666 | |
| 0 | 3 | male | 55.5 | 0 | 0.071074 | |
| 0 | 3 | male | 28 | 0 | 0.099323 | |
| 0 | 3 | female | 2 | 1 | 0.630698 | |
| 1 | 3 | female | 28 | 1 | 0.547262 | |
| 0 | 3 | male | 14 | 0 | 0.117255 | |
| 1 | 3 | male | 32 | 0 | 0.094666 | |
| 0 | 2 | male | 24 | 0 | 0.24768 | |
| 1 | 1 | female | 41 | 1 | 0.882931 | |
| 1 | 3 | male | 16 | 0 | 0.114532 | |
| 0 | 1 | male | 28 | 0 | 0.449892 | |
| 1 | 3 | male | 21 | 0 | 0.107962 | |
| 0 | 3 | male | 8 | 0 | 0.125766 | |
| 0 | 3 | male | 45 | 0 | 0.080858 | |
| 1 | 1 | male | 11 | 1 | 0.506216 | |
| 1 | 3 | female | 36 | 1 | 0.520807 | |
| 0 | 3 | male | 16 | 0 | 0.114532 | |
| 0 | 3 | male | 28 | 0 | 0.099323 | |
| 1 | 2 | female | 19 | 1 | 0.794101 | |
| 0 | 3 | male | 28 | 0 | 0.099323 | |
| Flight 4 | 1 | female | 22 | 0 | 0.095492 | |

To view, select a column in the table.

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

Kristen Chian 70

評估模型好壞

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|--|---|
| 預測 YES | True Positive (TP) False Negative (FN) Type II Error | False Positive (FP) Type I Error |
| 預測 NO | | True Negative (TN) |

評估模型好壞

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total N}}$$

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|--|-------------------------------------|
| 預測 YES | True Positive (TP) False Negative (FN) Type II Error | False Positive (FP) Type I Error |
| 預測 NO | | True Negative (TN) |

評估模型好壞

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|--------------------------------------|---------------------|
| 預測 YES | True Positive (TP) Type I Error | False Positive (FP) |
| 預測 NO | False Negative (FN) Type II Error | True Negative (TN) |

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$



評估模型好壞

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|--------------------------------------|-------------------------------------|
| 預測 YES | True Positive (TP) | False Positive (FP) Type I Error |
| 預測 NO | False Negative (FN) Type II Error | True Negative (TN) |

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



評估模型好壞

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|---|---------------------|
| 預測 YES | True Positive (TP) <small>Type I Error</small> | False Positive (FP) |
| 預測 NO | False Negative (FN) <small>Type II Error</small> | True Negative (TN) |

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

評估模型好壞

| Confusion Matrix | 實際 YES | 實際 NO |
|------------------|--------------------------------------|-------------------------------------|
| 預測 YES | True Positive (TP) | False Positive (FP) Type I Error |
| 預測 NO | False Negative (FN) Type II Error | True Negative (TN) |

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total N}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

評估模型好壞

Microsoft Azure Machine Learning Studio

Search experiment items

Saved Datasets

- My Datasets
 - Titanic Dataset
 - Titanic Dataset - Copy
 - Titanic Dataset - Copy (2)
 - Titanic_Test.csv
 - Titanic_Train.csv
- Samples
 - Adult Census Income Bl...
 - 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...
 - Energy Efficiency Regres...
 - Flight Delays Data
 - Flight on-time performa...
 - Forest fires data

→ 從 Module 選擇, Machine Learning 中 Evaluate 的 [Evaluate Model]

```
graph TD; A[Clean Missing Data] --> B[Clean Missing Data]; B --> C[Two-Class Logistic Regression]; C --> D[Split Data]; D --> E[Train Model]; E --> F[Score Model]; F --> G[Evaluate Model];
```

STATUS DATA... None

Prior Run

Summary

Description

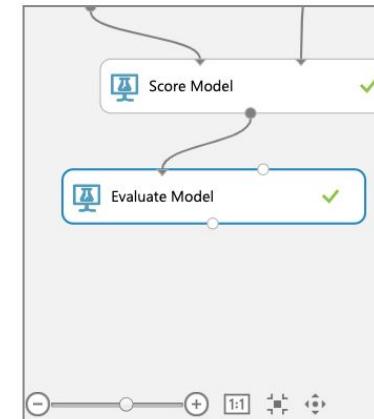
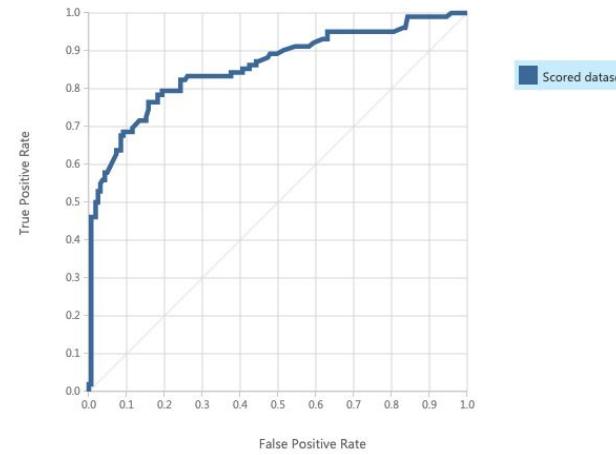
Quick Help

KristenChen 77

評估模型好壞

Kaggle Titanic Input3 > Evaluate Model > Evaluation results

ROC PRECISION/RECALL LIFT



| True Positive | False Negative | Accuracy | Precision | Threshold | AUC | | |
|----------------|----------------|----------|-----------|-----------|-------|--|--|
| 70 | 32 | 0.820 | 0.814 | 0.5 | 0.858 | | |
| False Positive | True Negative | Recall | F1 Score | | | | |
| 16 | 149 | 0.686 | 0.745 | | | | |
| Positive Label | Negative Label | | | | | | |
| 1 | 0 | | | | | | |

| Score Bin | Positive Examples | Negative Examples | Fraction Above Threshold | Accuracy | F1 Score | Precision | Recall | Negative Precision | Negative Recall | Cumulative AUC |
|---------------|-------------------|-------------------|--------------------------|----------|----------|-----------|--------|--------------------|-----------------|----------------|
| (0.900,1.000] | 3 | 1 | 0.015 | 0.625 | 0.057 | 0.750 | 0.029 | 0.624 | 0.994 | 0.000 |

發布 Web Service

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3

Machine Learning

- Evaluate
 - Cross Validate Model
 - Evaluate Model
 - Evaluate Recommender
- Statistical Functions
- Evaluate Probability Function

Properties Project

Experiment Properties

- START TIME 5/17/2019 ...
- END TIME 5/17/2019 ...
- STATUS CODE Finished
- STATUS DATA... None

Prior Run

Summary

Description

Mini Map

1 Predictive Web Service [Recommended]

Retraining Web Service

NEW

SAVE AS

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

https://studio.azureml.net/Home/ViewWorkspaceCached/2f8820c3214848a59ba3304246846ead#

```
graph TD; EM[Edit Metadata] --> C1[Clean Missing Data]; C1 --> C2[Clean Missing Data]; C2 --> SD[Split Data]; SD --> LR[Two-Class Logistic Regression]; SD --> TM[Train Model]; TM --> SM[Score Model]; SM --> EM2[Evaluate Model];
```

openChian 79

發布 Web Service

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3 [Predictive Exp.]

In draft. Draft saved at 午后3:53:41

Properties Project

Experiment Properties

- START TIME 5/17/2019 ...
- END TIME 5/17/2019 ...
- STATUS CODE InDraft
- STATUS DATA... None

Summary

Description

Mini Map

Creating predictive experiment

2

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

```
graph TD; WSInput[Web service input] --> ImportData[Import Data]; ImportData --> SelectColumns[Select Columns in Dataset]; SelectColumns --> EditMetadata[Edit Metadata]; EditMetadata --> Clean1[Kaggle Titanic Input3 (Clean...)]; Clean1 --> Clean2[Kaggle Titanic Input3 (Clean...)]; Clean2 --> ApplyT1[Apply Transformation]; Clean2 --> ApplyT2[Apply Transformation]; ApplyT1 --> ScoreModel[Score Model]; ApplyT2 --> ScoreModel; ScoreModel --> WSOutput[Web service output]
```

Kristen-Free-Workspace

DETAILS i CLOSE x

KristenChian 80

發布 Web Service

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3 [Predictive Exp.]

Training experiment Predictive experiment

Import Data Web service input

Select Columns in Dataset

Edit Metadata

Kaggle Titanic Input3 [Clean...]

Kaggle Titanic Input3 [Clean...]

Apply Transformation

Kaggle Titanic Input3 [traine...]

Score Model

Web service output

注意
我們 Model 是挑三個變數去預測而已,
所以 Web Input 擺這裡會有問題

Properties Project

Experiment Properties

START TIME 5/17/2019 ...
END TIME 5/17/2019 ...
STATUS CODE Finished
STATUS DATA... None

Go to web service

Summary

Description

Quick Help

Kristen-Free-Workspace

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

DEPLOY WEB SERVICE

PUBLISH TO GALLERY

Kristen Chian 81

發布 Web Service

Kaggle Titanic Input3 [Predictive Exp.]

Properties Project

Experiment Properties

START TIME 5/17/2019 ...
END TIME 5/17/2019 ...
STATUS CODE Finished

Import Data → Select Columns in Dataset → Edit Metadata → Kaggle Titanic Input3 (Clean...) → Kaggle Titanic Input3 (Clean...) → Kaggle Titanic Input3 (train...) → Select Columns in Dataset → Web service input → Score Model → Web service output

3

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

Mini Map

Search experiment items

Saved Datasets

My Datasets

- Titanic Dataset
- Titanic Dataset - Copy
- Titanic Dataset - Copy (2)
- 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 5...
- Energy Efficiency Regres...
- Flight Delays Data
- Flight on-time performa...
- Forest fires data

Select columns

BY NAME WITH RULES

AVAILABLE COLUMNS All Types search columns

Survived

SELECTED COLUMNS All Types search columns

Pclass
Sex
Age

1 columns available 3 columns selected

Quick Help

Kristen-Free-Workspace

Kirilen Chiaro 82

發布 Web Service

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3 [Predictive Exp.]

Properties Project Experiment Properties

START TIME 5/17/2019 ...
END TIME 5/17/2019 ...
STATUS CODE Finished
STATUS DATA... None

Go to web service

Import Data → Select Columns in Dataset → Edit Metadata → Kaggle Titanic Input3 [Clean...]

Select Columns in Dataset → Score Model → Select Columns in Dataset → Web service output

4

Select columns

BY NAME WITH RULES

AVAILABLE COLUMNS All Types search columns

Pclass Sex Age Scored Probabilities

SELECTED COLUMNS All Types search columns

Scored Labels

4 columns available 1 columns selected

Mini Map

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

Kristen Chau 83

Kaggle Titanic Input3 [Predictive Exp.] test returned ["1","female","22","1","0.906619071960449"]...

發布 Web Service

Microsoft Azure Machine Learning Studio

Kaggle Titanic Input3 [Predictive Exp.]

Import Data → Select Columns in Dataset → Edit Metadata → Kaggle Titanic Input3 [Clean...]

Kaggle Titanic Input3 [Clean...] → Apply Transformation → Kaggle Titanic Input3 [Train...]

Kaggle Titanic Input3 [Train...] → Select Columns in Dataset → Score Model → Select Columns in Dataset → Web service output

Web service input → Apply Transformation → Kaggle Titanic Input3 [Train...]

Properties Project Experiment Properties Summary Description

Kristen-Free-Workspace

5

Kristen Chan 84

Quick Help

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

發布 Web Service

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

kaggle titanic input3 [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

oHn

Default Endpoint

API HELP PAGE

REQUEST/RESPONSE

BATCH EXECUTION

TEST

6 Test Test: Preview

APP

LAST UPDATED

Excel 2013 or later | Excel 2010 or earlier workbook

Excel 2013 or later workbook

5/17/2019 4:05:53 PM

5/17/2019 4:05:53 PM

選 Test 做測試

NEW

DELETE

Chiau 85

Advanced Version -- Python

Enter data to predict 測試

PCLASS

1

SEX

female

AGE

22

Azure ML Studio with Python Notebook

建立一個 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 highlighted with a pink border), DATASETS, TRAINED MODELS, and SETTINGS. At the bottom left of the sidebar is a 'NEW' button with a plus sign. The main area is titled 'notebooks' with a 'preview' link. It contains a table header with columns: NAME, LANGUAGE, LAST MODIFIED, and PROJECT. Below the header, it says 'No notebooks found'. At the bottom of the main area are three buttons: DELETE, RENAME, and ADD TO PROJECT.

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

notebooks [preview](#)

| NAME | LANGUAGE | LAST MODIFIED | PROJECT |
|--------------------|----------|---------------|---------|
| No notebooks found | | | |

PROJECTS

EXPERIMENTS

WEB SERVICES

NOTEBOOKS

DATASETS

TRAINED MODELS

SETTINGS

NEW

DELETE

RENAME

ADD TO PROJECT

建立一個 Python3 Notebook

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace    

PROJECTS notebooks 

| NAME | LANGUAGE | LAST MODIFIED | PROJECT |
|---|----------|---------------|---------|
|  X | | | |

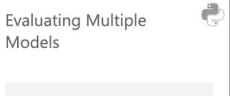
NEW

 DATASET  MODULE  PROJECT  EXPERIMENT  NOTEBOOK  PREVIEW

 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  Access Azure ML Experiment Data  Variable Selection in Azure ML Jupyter Notebook  GBM in Azure ML Jupyter Notebook  Evaluating Multiple Models

KristenChian 89

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

Krispen Chian 90

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 user profile icons.

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

DELETED 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())

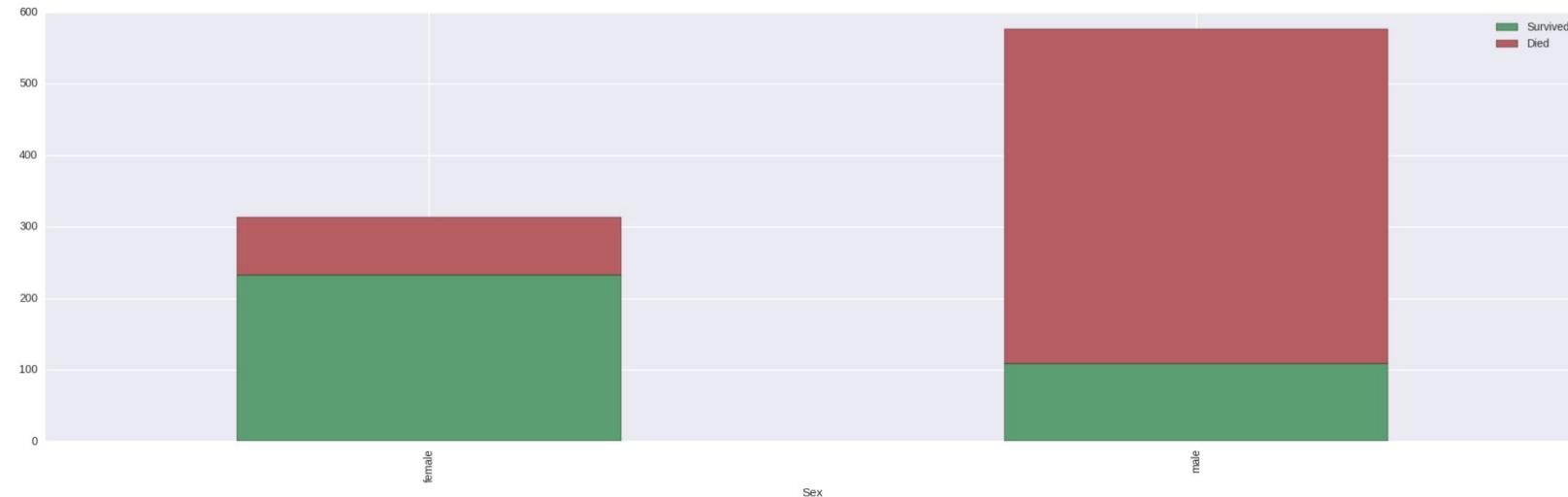
Sex VS. Survived (count)

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] :

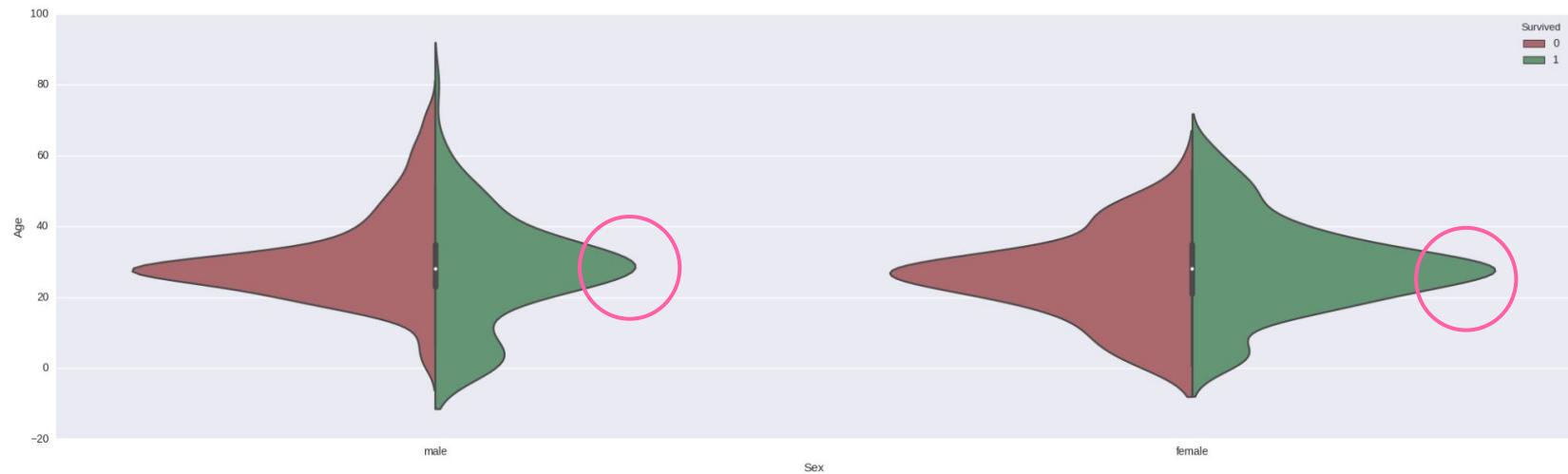
```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Sex', y = 'Age',
                hue = 'Survived', data = data_train,
                split = True,
                palette = {0: "r", 1: "g"})
);
```



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"})
);
```

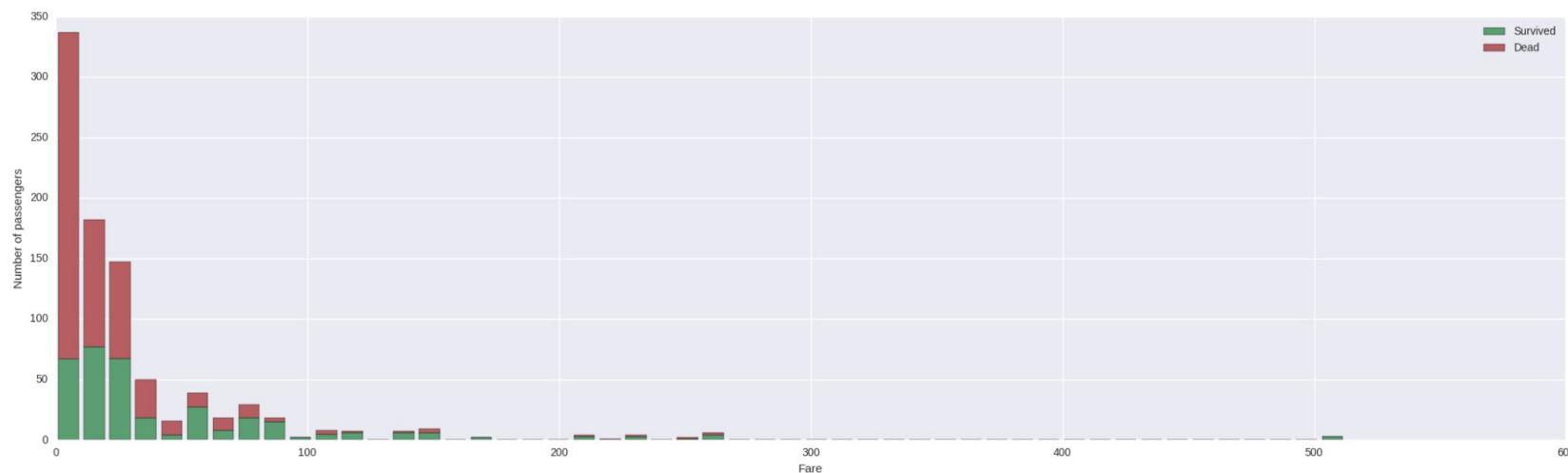


KristenChian99

Fare VS. Survived

In [10]:

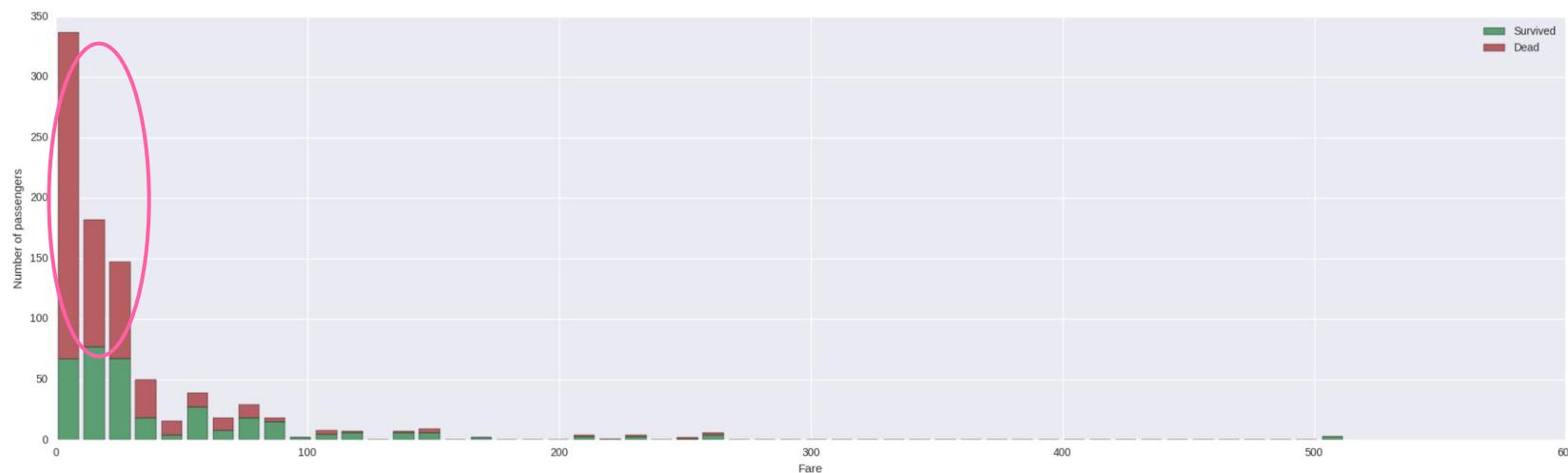
```
figure = plt.figure(figsize = (25, 7))
plt.hist([data_train[data_train['Survived'] == 1]['Fare'], data_train[data_train['Survived'] == 0]['Fare']],
         stacked = True, color = ['g','r'],
         bins = 50, label = ['Survived', 'Dead'])
plt.xlabel('Fare')
plt.ylabel('Number of passengers')
plt.legend();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



Fare VS. Survived

In [10]:

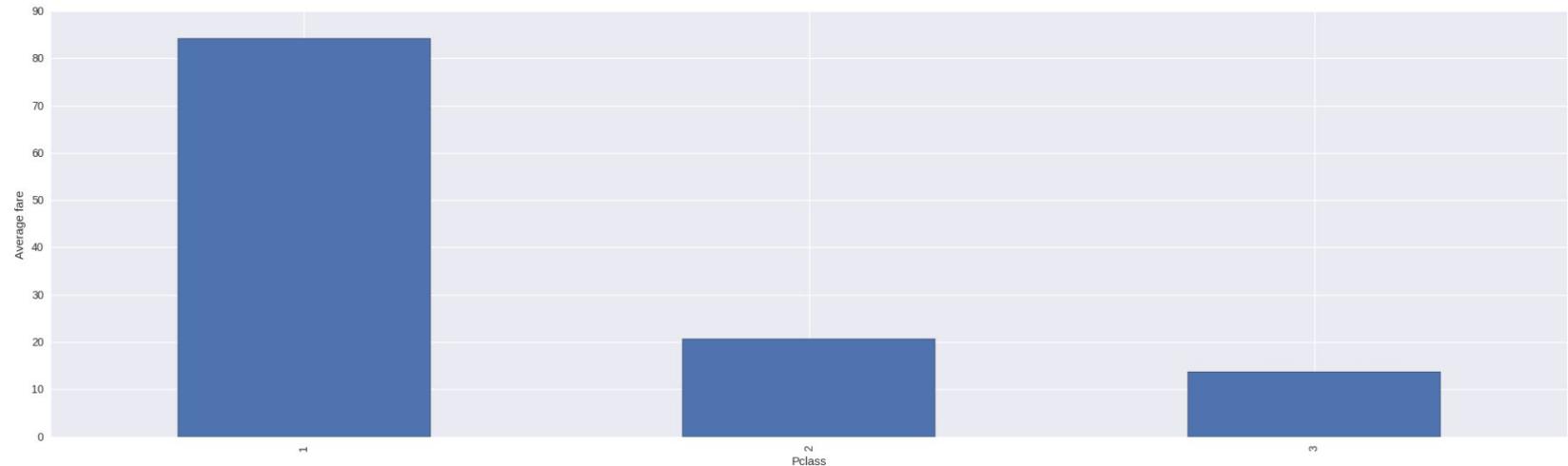
```
figure = plt.figure(figsize = (25, 7))
plt.hist([data_train[data_train['Survived'] == 1]['Fare'], data_train[data_train['Survived'] == 0]['Fare']],
         stacked = True, color = ['g','r'],
         bins = 50, label = ['Survived', 'Dead'])
plt.xlabel('Fare')
plt.ylabel('Number of passengers')
plt.legend();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



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] :

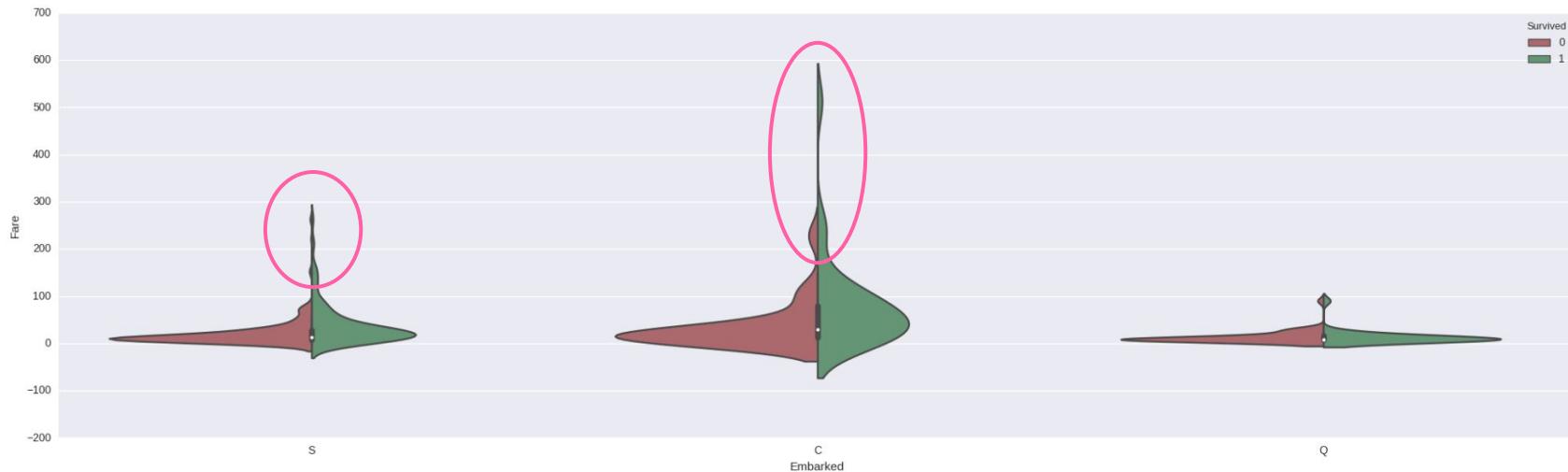
```
fig = plt.figure(figsize = (25, 7))
sns.violinplot(x = 'Embarked', y = 'Fare', hue = 'Survived', data = data_train, split = True,
                 palette = {0: "r", 1: "g"});
```



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  
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
```

$$891 + 418 = 1309$$

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

Python Script

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

Saved Datasets

Data Format Converters

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

Inscribing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

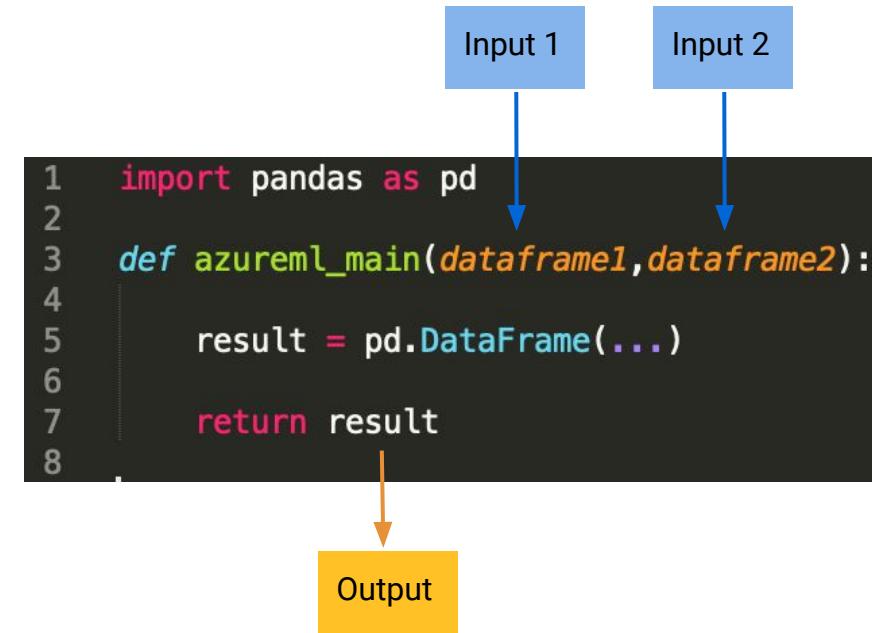
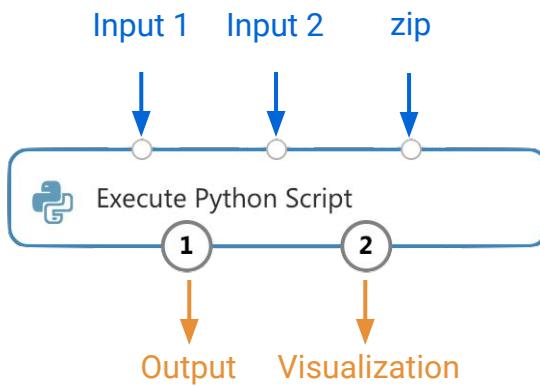
RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

Kristen Chia 124

介紹 Excute Python Script



Python Script -- Titanic Feature Engineering

Screenshot of a GitHub repository page for a Python script named "Titanic_FeatureEngineering.py".

The repository URL is https://github.com/kristenchan/Sharing/blob/master/Kaggle_Titanic/Titanic_FeatureEngineering.py.

The repository details show it was last updated 53c658b a minute ago by user [kristenchan](#). It has 0 forks and 0 contributors.

The code editor shows the first 22 lines of the 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",
22        "Mme": "Mrs",
```



KristenChan 126

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 summary 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

NEW

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

Kristen Chau 127

Excute Python Script

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

Kaggle Titanic Advance In draft Saving... Search experiment items

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",
```

Execute Python Script

1 2

Quick Help

Execute a Python script from an Azure Machine Learning experiment.

Execute Python Script

1 2

Quick Help

Execute 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]

Excute Python Script

Microsoft Azure Machine Learning Studio

Kaggle Titanic Advance

Finished running ✓ Properties Project

rows 891 columns 69

view as

| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Title | Title_Master | Title_Miss | Title |
|-------------|----------|--------|---|--------|-----|-------|-------|-----------------|---------|-------|----------|--------|--------------|------------|-------|
| 1 | 0 | 3 | Braund, Mr. Owen Harris | Male | 22 | 1 | 0 | A/5 21171 | 7.25 | M | S | Mr | 0 | 0 | 1 |
| 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | Female | 38 | 1 | 0 | PC 17535 | 71.2833 | C | C | Mrs | 0 | 0 | 0 |
| 3 | 1 | 3 | Heikkinen, Miss. Laina | Female | 26 | 0 | 0 | STON/O2 3101283 | 7.925 | M | S | Miss | 0 | 1 | 0 |
| 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | Female | 35 | 1 | 0 | 3101282 | 53.1 | C | S | Mrs | 0 | 0 | 0 |
| 5 | 0 | 3 | Allan, Mr. William Henry | Male | 35 | 0 | 0 | 3101281 | 8.05 | M | S | Mr | 0 | 0 | 1 |
| 6 | 0 | 3 | Moran, Mr. James | Male | 26 | 0 | 0 | 3101280 | 8.4583 | M | Q | Mr | 0 | 0 | 1 |
| 7 | 0 | 1 | McCarthy, Mr. Timothy J | Male | 54 | 0 | 0 | 3101279 | 51.8625 | E | S | Mr | 0 | 0 | 1 |
| 8 | 0 | 3 | Palsson, Master. Gosta Leonard | Male | 2 | 3 | 1 | 3101278 | 21.075 | M | S | Master | 1 | 0 | 0 |
| 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | Female | 27 | 0 | 2 | 3101277 | 11.1333 | M | S | Mrs | 0 | 0 | 0 |
| 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | Female | 14 | 1 | 0 | 3101276 | 30.0708 | M | C | Mrs | 0 | 0 | 0 |
| 11 | 1 | 3 | Sandstrom, Miss. Marguerite Rut | Female | 4 | 1 | 1 | 3101275 | 16.7 | G | S | Miss | 0 | 1 | 0 |
| 12 | 1 | 1 | Bonnell, Miss. Elizabeth | Female | 58 | 0 | 0 | 3101274 | 26.55 | C | S | Miss | 0 | 1 | 0 |

To view, select a column in the table.

Statistics

Visualizations

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

NEW Kristen Chaw 129

Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kaggle Titanic Advance

Titanic_Train.csv

Execute Python Script

Select Columns in Dataset

Saved Datasets

Samples

Restaurant customer data

Restaurant feature data

Data Transformation

Manipulation

- Join Data
- Select Columns in Dataset...
- Select Columns Transfor...

Feature Selection

- Filter Based Feature Select...
- Fisher Linear Discriminant ...
- Permutation Feature Import...

Statistical Functions

- Compute Elementary Statisti...

Text Analytics

- Extract N-Gram Features fr...

Mini Map

PassengerId

Pclass

Name

Ticket

Cabin

Embarked

Title

Survived

Sex

Age

Title_Master

Title_Miss

Title_Mr

Title_Mrs

Title_Officer

Title_Royalty

Embarked_C

Embarked_Q

Embarked_S

Cabin_A

Cabin_B

Cabin_C

PassengerId

Pclass

Name

Title

Embarked

Cabin

Ticket

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

NEW

篩選欄位

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

Select columns

BY NAME
WITH RULES

AVAILABLE COLUMNS

All Types | search columns

PassengerId
Pclass
Name
Ticket
Cabin
Embarked
Title

SELECTED COLUMNS

All Types | search columns

Survived
Sex
Age
Title_Master
Title_Miss
Title_Mr
Title_Mrs
Title_Officer
Title_Royalty
Embarked_C
Embarked_Q
Embarked_S
Cabin_A
Cabin_B
Cabin_C



留下

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

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.

(more help...)

Kristen Chan 130

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.

Mini Map

Execute Python Script

Select Columns in Dataset

Two-Class Logistic Regression

Split Data

Train Model

Score Model

Evaluate Model

清理資料

Run History Save 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]; C --> E[Split Data]; D --> F[Train Model]; E --> F; F --> G[Score Model]; G --> H[Evaluate Model]
```

Advanced Version -- Python

Microsoft Azure Machine Learning Studio Kristen-Free-Workspace ?

Kaggle Titanic Advance

Search experiment items

Saved Datasets Trained Models 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 Deprecated

Properties Project

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

Predictive Web Service [Recommended] Retraining Web Service

Titanic_Train.csv

Execute Python Script

Select Columns in Dataset

Two-Class Logistic Regression

Split Data

Train Model

Score Model

Evaluate Model

Diagram:

```
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]
```

Bottom Bar:

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

KrisjenChian 132

Advanced Version -- Python

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

[Web Service] Step 1 Training experiment Predictive experiment 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[Web service input] --> B; B --> C[Select Columns in Dataset]; C --> D[Score Model]; D --> E[Web service output]; F[Kaggle Titanic Advance [train...]] --> C;
```

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

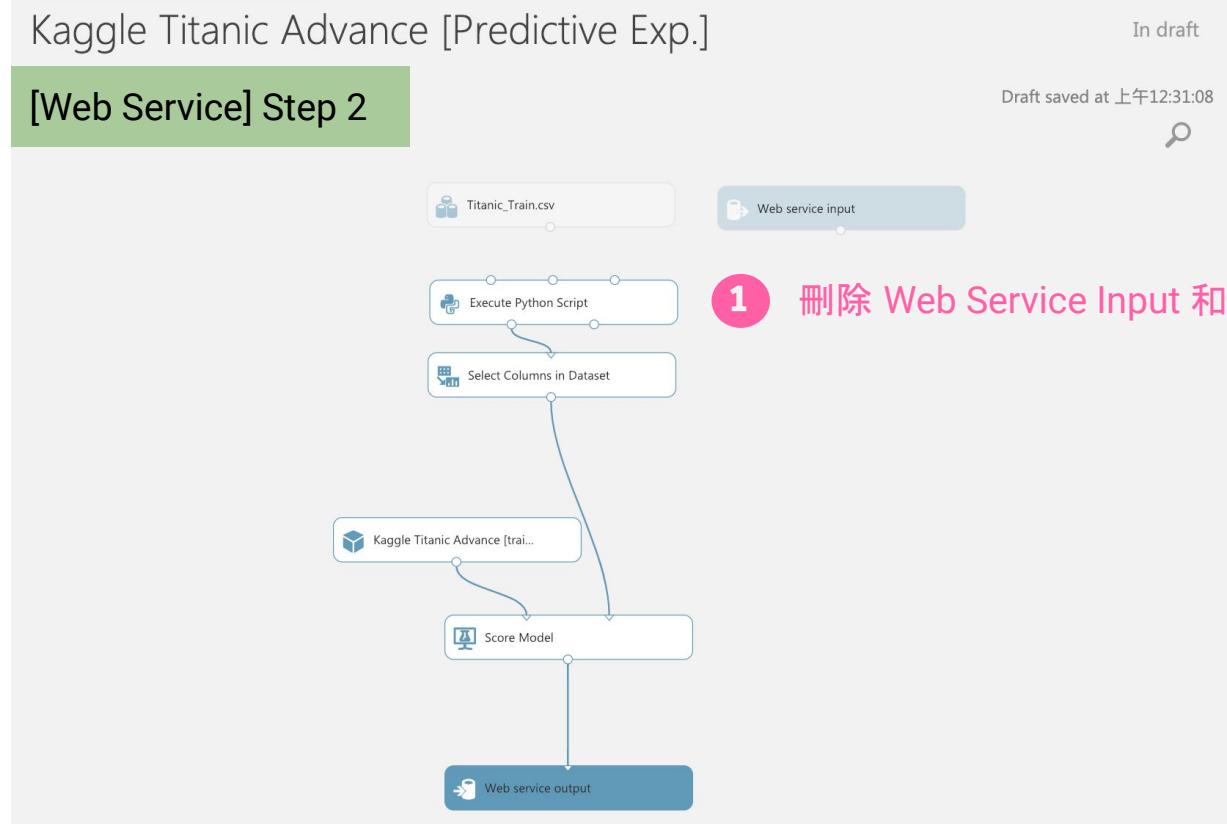
Enter the detailed description for your experiment.

Quick Help

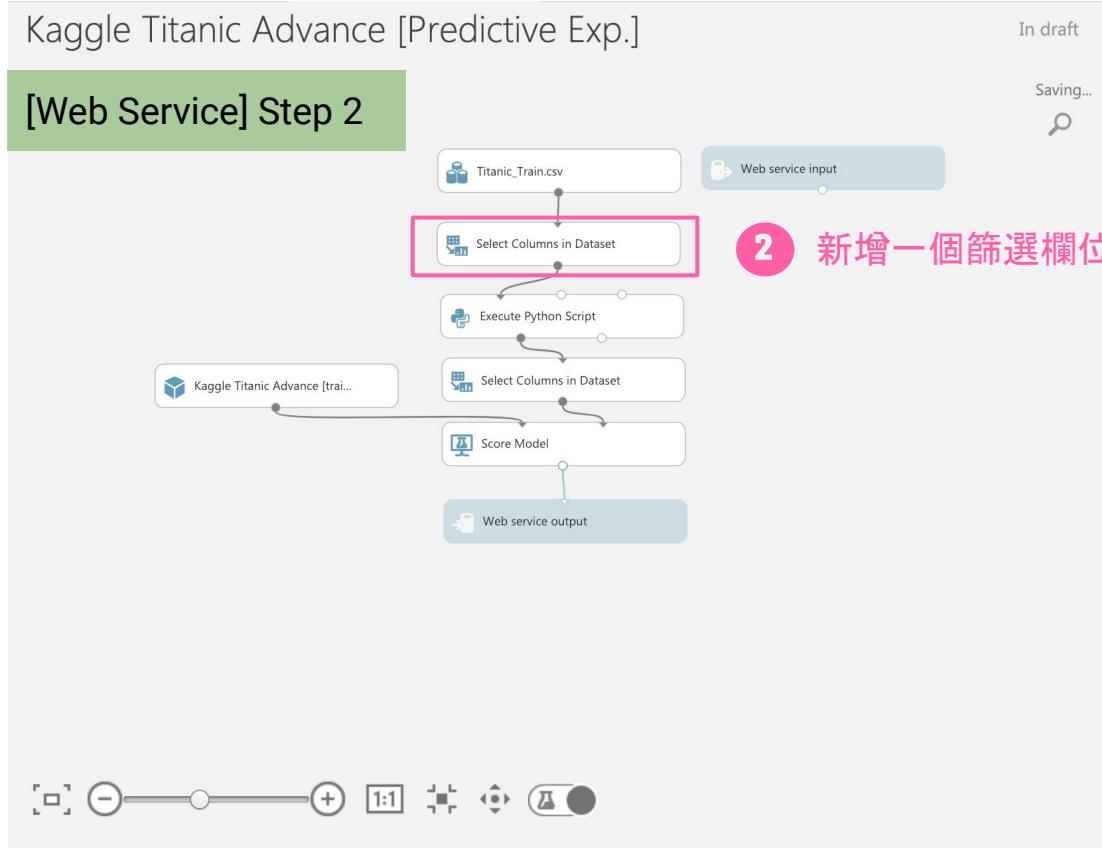
NEW RUN HISTORY SAVE DISCARD CHANGES RUN DEPLOY WEB SERVICE PUBLISH TO GALLERY

KrisenChian 133

Advanced Version -- Python



Advanced Version -- Python



Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

[Web Service] Step 2

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

Selected columns:

Column names:

Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked

Launch column selector

3 選擇 Python Script 中會處理到的變數

AVAILABLE COLUMNS

BY NAME

WITH RULES

All Types search columns

PassengerId
Survived

SELECTED COLUMNS

All Types search columns

Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked

2 columns available

10 columns selected

[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

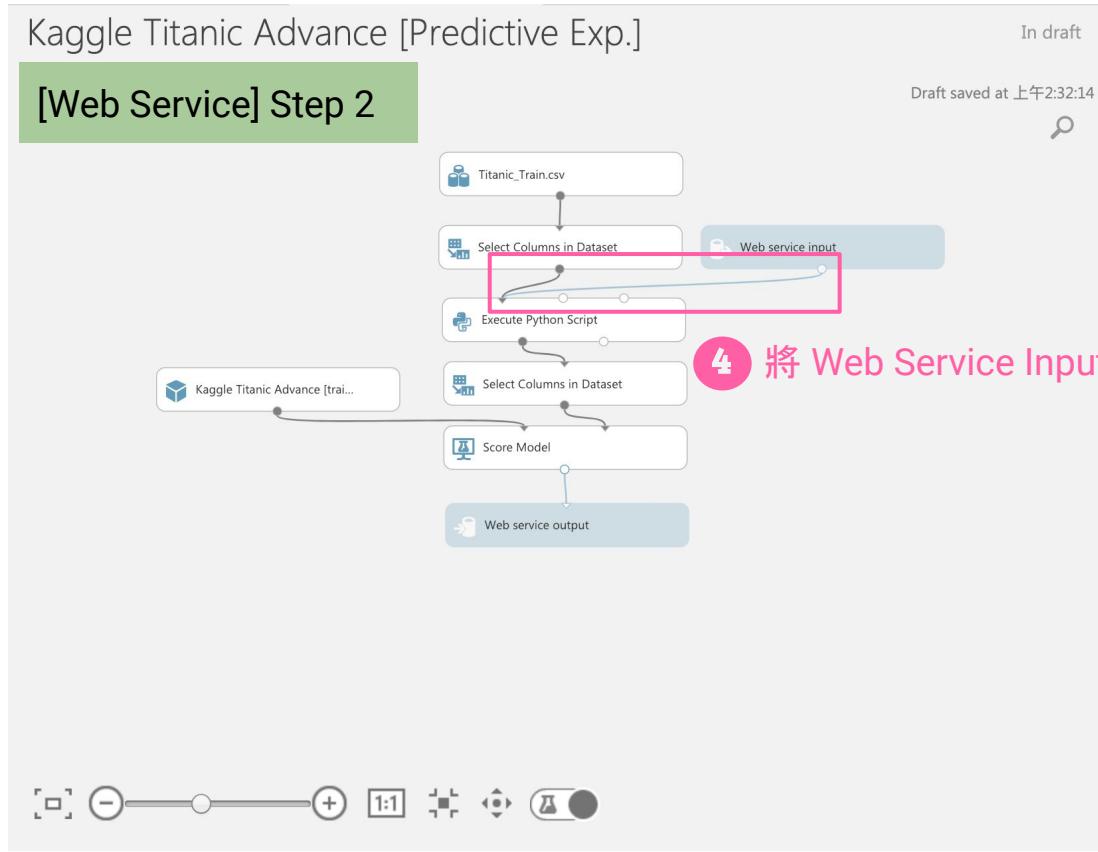
Quick Help

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

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

Yifan Chen 136

Advanced Version -- Python



Advanced Version -- Python

Microsoft Azure Machine Learning Studio Kristen-Free-Workspace

[Web Service] Step 2 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

Training experiment Predictive experiment

Properties Project

▪ Select Columns in Dataset

Select columns

Selected columns:
Column names:
Survived,Sex,Age,Title,Maste...

Launch column selector

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.

5 將原本的 Select Columns in Dataset 中的 Survived 刪除

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

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

Advanced Version -- Python

Microsoft Azure Machine Learning Studio

Kristen-Free-Workspace

[Web Service] Step 2

Predictive experiment

Kaggle Titanic Advance [Predictive Exn]

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...)

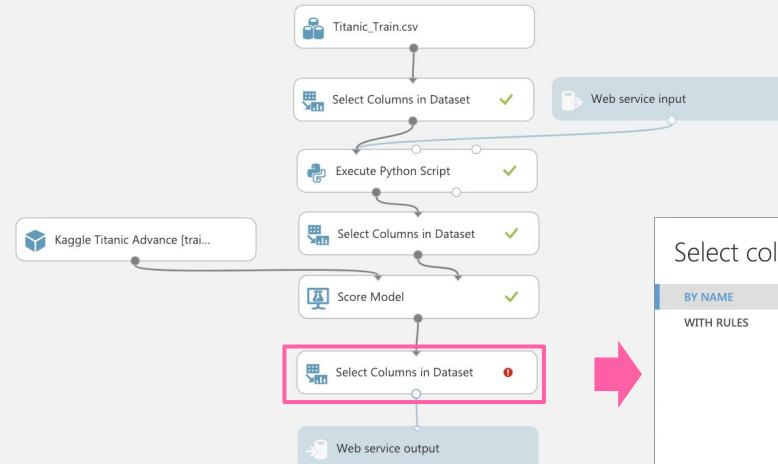
Run History Save Save As Discard Changes Run Deploy Web Service Publish to Gallery

KristenChian 139

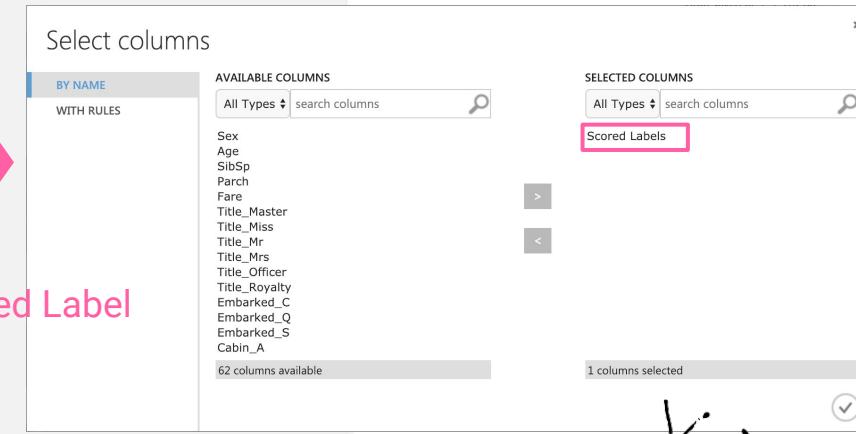
Advanced Version -- Python

Kaggle Titanic Advance [Predictive Exp.]

[Web Service] Step 2



- 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
Web service input
Execute Python Script
Select Columns in Dataset
Kaggle Titanic Advance [trai...
Score Model
Select Columns in Dataset
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[Select Columns in Dataset]; D --> E[Score Model]; E --> F[Select Columns in Dataset]; F --> G[Web service output];
```

KristenChian 141

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

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES 1:1 🔍 ⚙️ 🔍

KristenChiaro 142

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 Test [Test preview](#)

BATCH EXECUTION Test [Test preview](#)

Excel 2013 or later | Excel 2010 or earlier workbook 2/26/2019 2:43:22 AM

Excel 2013 or later workbook 2/26/2019 2:43:22 AM

 NEW 

[Web Service] Step 4



143

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 [preview](#)

BATCH EXECUTION TEST [preview](#)

選 Test 做測試

[Web Service] Step 4

KristenChian 144

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+C0W9Y4JDoeYeg==

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 |
|--|---------------------|--------------------------------|----------------------|
| Excel 2013 or later workbook | | | 2/27/2019 1:35:03 AM |
| Excel 2013 or later workbook | | | 2/27/2019 1:35:03 AM |

成功 !!

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

DETAILS i CLOSE X

DELETE

KrisfenChian 146

Note

Reference

- Kaggle Titanic
 - 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>