

資料科學研究院 挖掘出大數據的隱藏寶藏- Azure Machine Learning

Allan Yiin

asiaMiner

CURVE UP THE
FUTURE



Agenda

- 微軟雲端大數據分析架構
- Azure Machine Learning架構介紹
- 主要功能介紹
- 案例：建置回應模型
- 案例：客戶區隔模型
- 整合R語言
- 整合Power BI for Office 365

A woman with brown hair in a ponytail, wearing a black and white striped shirt, is pointing at a tablet. A man with grey hair and glasses, wearing a dark suit jacket over a light blue shirt, is looking at the tablet. The background is a plain, light-colored wall.

雲端大數據分析 架構

大數據時代的到來

ExaByte
(10E18)

PetaByte
(10E15)

TeraByte
(10E12)

GigaByte
(10E9)

大數據



快速增長

多變

多樣性



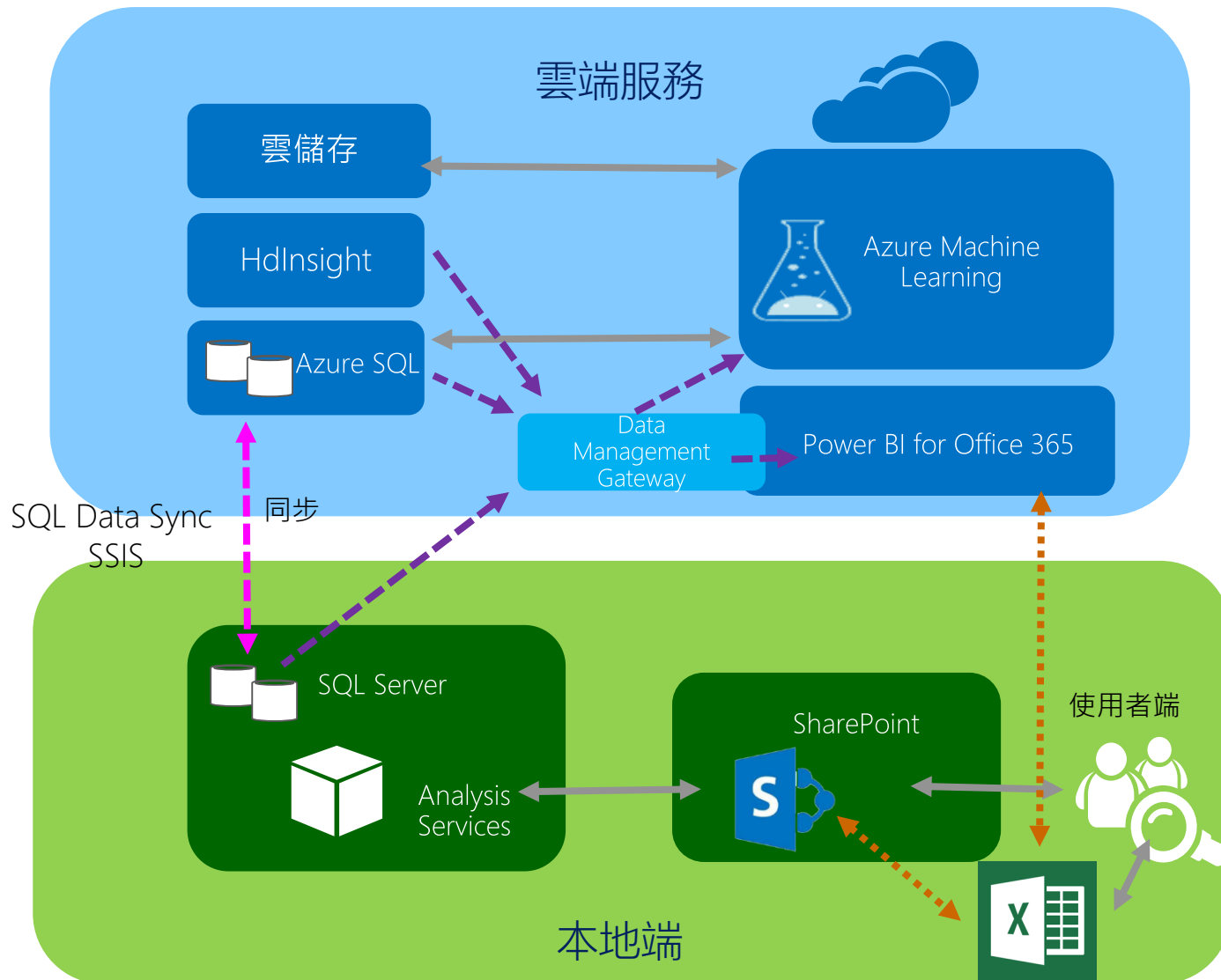
1980
190,000\$

1990
9,000\$

2000
15\$

2010
0.07\$

雲端與本地端的分析架構



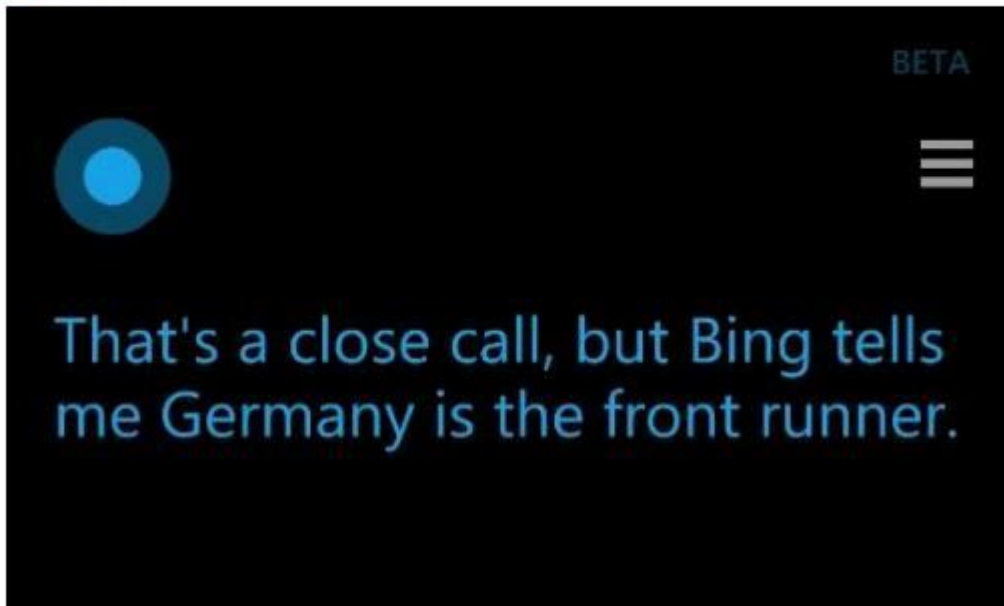
有了大數據之後 如何從中產生效益...?



Cortana @cortanaquotes · 16分

Will Germany win against Argentina?

[查看翻譯](#)



微軟Cortana世界盃足球賽16場正確預測15場比賽...
擁有數據，就能預測這世界的未來...

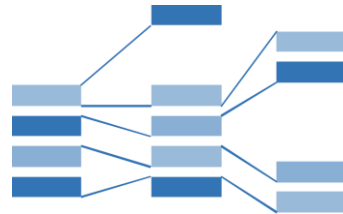
A woman with brown hair in a ponytail, wearing a black and white striped shirt, is pointing at a tablet. A man with grey hair and glasses, wearing a dark suit and a light blue shirt, is looking at the tablet. The background is a plain, light-colored wall.

Azure Machine Learning

Machine Learning 在做些甚麼...?

透過演算法協助我們從龐大的數據中找出有意義的規則

分類



分群



推估



Azure Machine Learning

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow diagram for a credit card credit model prediction. The workflow starts with a data source 'Creditcard Credit Model.csv', followed by 'Project Columns', 'Split' (labeled '切割訓練集與測試集'), 'Filter Based Feature Selection' (labeled '基於過濾的特徵選擇'), and 'Train Model' (labeled '訓練模型'). A 'Two-Class Decision Forest' (labeled '決策森林模型') is also shown as a model component. The workflow is marked as 'Finished running'. The right-hand side shows the 'Properties' panel for the experiment, including 'Experiment Properties' with details like 'START TIME', 'END TIME', 'STATUS CODE', and 'STATUS DETAILS'. The bottom of the interface features a navigation bar with icons for 'NEW', 'VIEW RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'REFRESH', 'CANCEL', 'RUN', and 'PUBLISH WEB SERVICE', along with a search bar for help content.

範例--信用評等預測模型

Microsoft Azure Machine Learning

Enter feedback here

AsiaMiner

Menu

Search experiment items

Saved Datasets

Data Format Conversions

Data Input and Output

Data Transformation

Feature Selection

R Language Modules

Regression

Filter Based Feature Selection

Train Model

Project Columns

Split

Project Columns

Two-Class Decision Forest

Train Model

Finished running

Properties

Experiment Properties

START TIME	10/9/2014 12:40:...
END TIME	10/9/2014 12:40:...
STATUS CODE	Finished
STATUS DETAILS	None

Go to web service

Disable upgrades

Prior Run

NEW

VIEW RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

REFRESH

CANCEL

RUN

PUBLISH WEB SERVICE

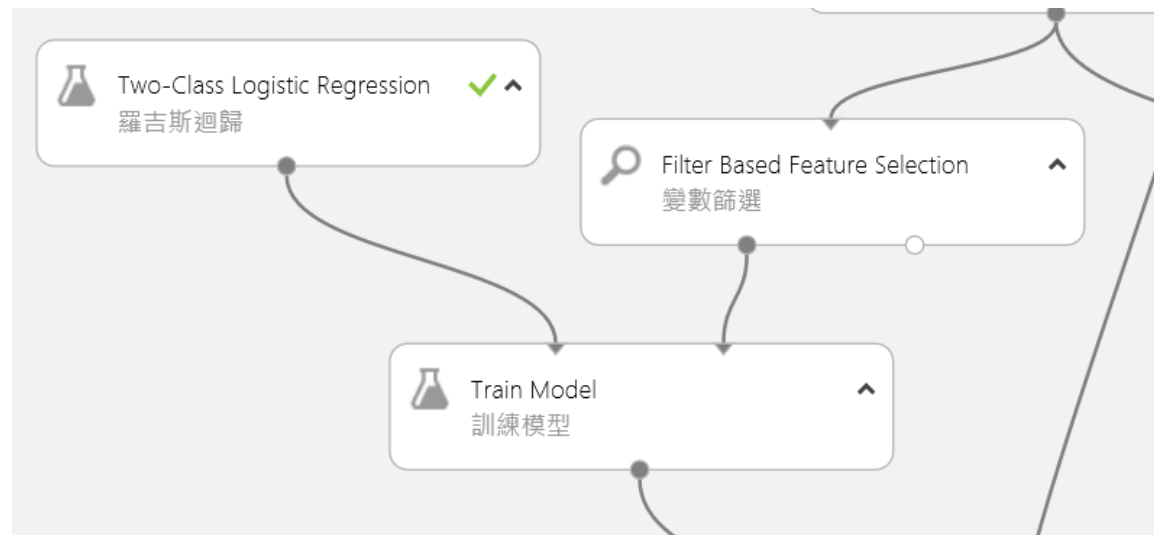
Search Help Content

透過瀏覽器存取，
將分析功能透過流
程圖的模式進行設
計...



















免費版本：
<https://studio.azureml.net/>

Azure Machine Learning 包括了...

- 資料源的存取(與Azure SQL Database, Azure Storage以及HDInsight深度整合)
- 資料清理(抽樣、樣本切割、計算欄位...)
- 敘述性統計
- 有效變數選取
- 建立預測模型
- 評估預測模型
- 應用預測模型



主要功能介紹

- ▶  Saved Datasets 
- ▶  Data Format Conversions 
- ▶  Data Input and Output 
- ▶  Data Transformation 
- ▶  Feature Selection 
- ▶  Machine Learning 
- ▶  R Language Modules 
- ▶  Statistical Functions 
- ▶  Text Analytics 

新增資料集或Experiment

Microsoft Azure Machine Learning | Home Studio

EXPERIMENTS

WEB SERVICES

SETTINGS

experiments

ALL EXPERIMENTS SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED
<input checked="" type="checkbox"/>	Cardif Winback Model	allan	Failed	9/2/2014 12:11:02 AM
<input type="checkbox"/>	POC--CEA Churn	raymondlinmission	Draft	11/5/2014 10:54:59 AM

NEW

DATASET

EXPERIMENT

New Experiment List

Search experiment templates

Microsoft Samples

- Blank Experiment
- Sample 1: Download dataset from...
- Sample 2: Dataset Processing and...
- Sample 3: Cross Validation for Bina...
- Sample 4: Cross Validation for Reg...
- Sample 5: Train, Test, Evaluate for...
- Sample 6: Train, Test, Evaluate for...
- Sample 7: Train, Test, Evaluate for...

上傳資料集

Upload a new dataset

Select the data to upload:

瀏覽...

This is the new version of an existing dataset

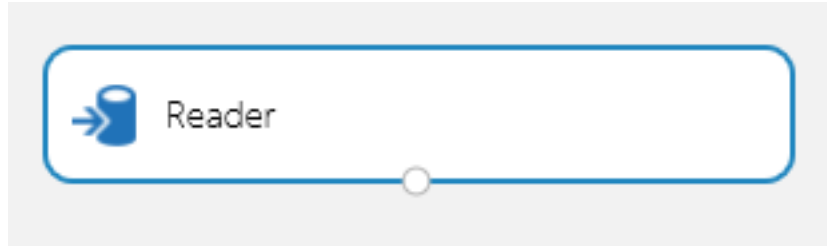
Enter a name for the new dataset:

Select a type for the new dataset:

- Select a dataset type...
- Generic CSV File with a header (.csv)
- Generic CSV File With no header (.nh.csv)
- Generic TSV File with a header (.tsv)
- Generic TSV File With no header (.nh.tsv)
- Plain Text (.txt)
- SvmLight File (.svmlight)
- Attribute Relation File Format (.arff)
- Zip File (.zip)
- R Object or Workspace (.RData)

包含自訂R Package
(Zip)以及RData整合

雲端資料讀取

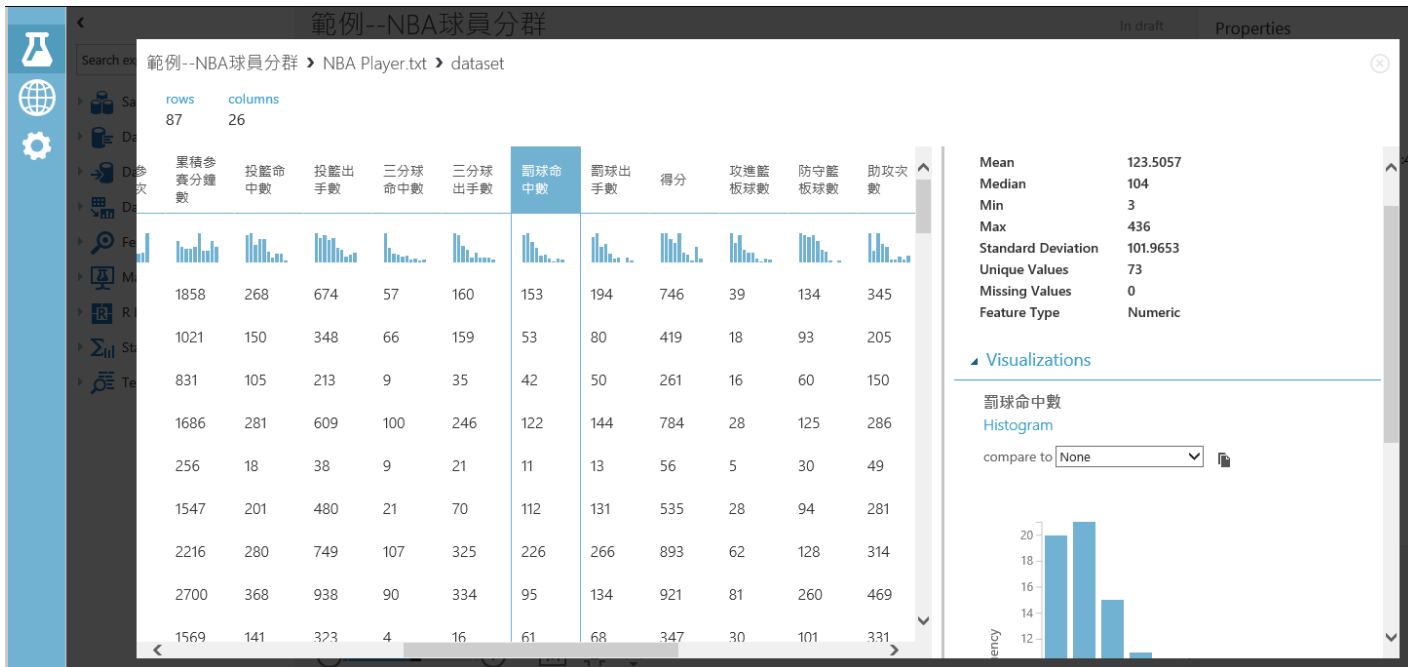
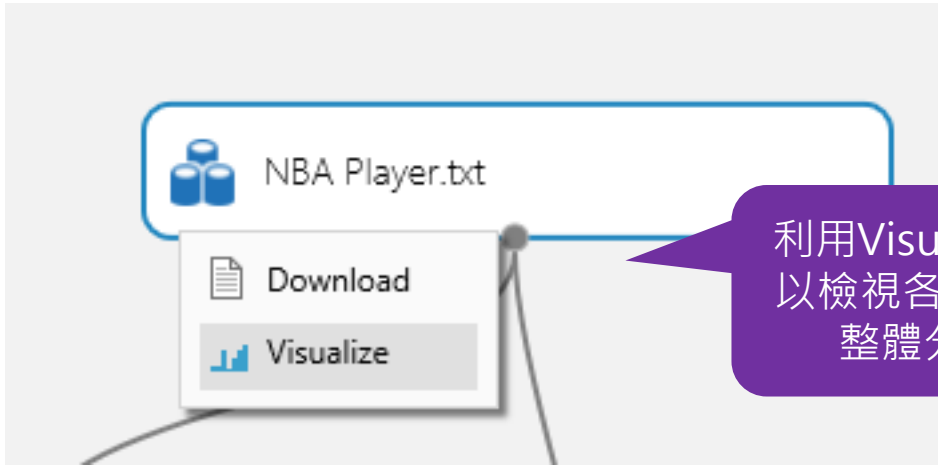


Properties

- Http
- SqlAzure
- AzureTable
- AzureBlobStorage
- HiveQuery
- PowerQuery
- Account

可以整合目前微軟各項雲端數據源技術

變數敘述性統計

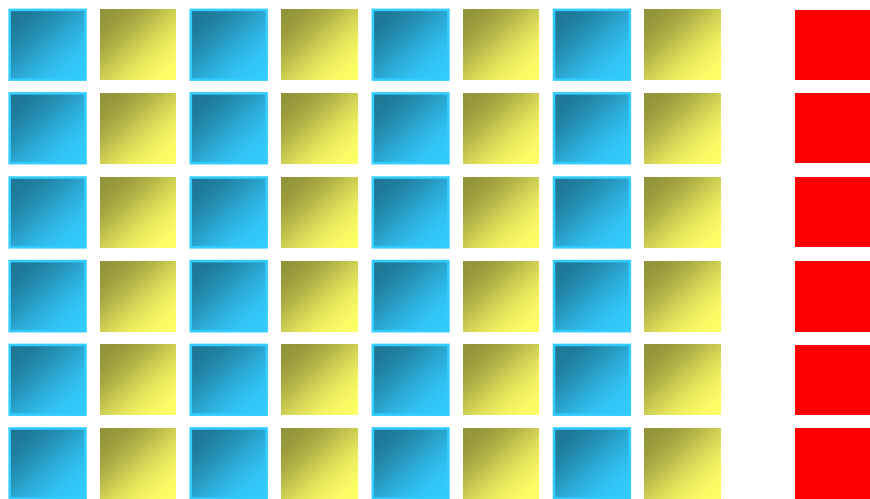


A woman with brown hair in a ponytail, wearing a black and white striped shirt, is pointing at a tablet. A man with grey hair and glasses, wearing a dark suit and a light blue shirt, is looking at the tablet. The background is a plain, light-colored wall.

建置 回應模型

什麼是預測？

案例

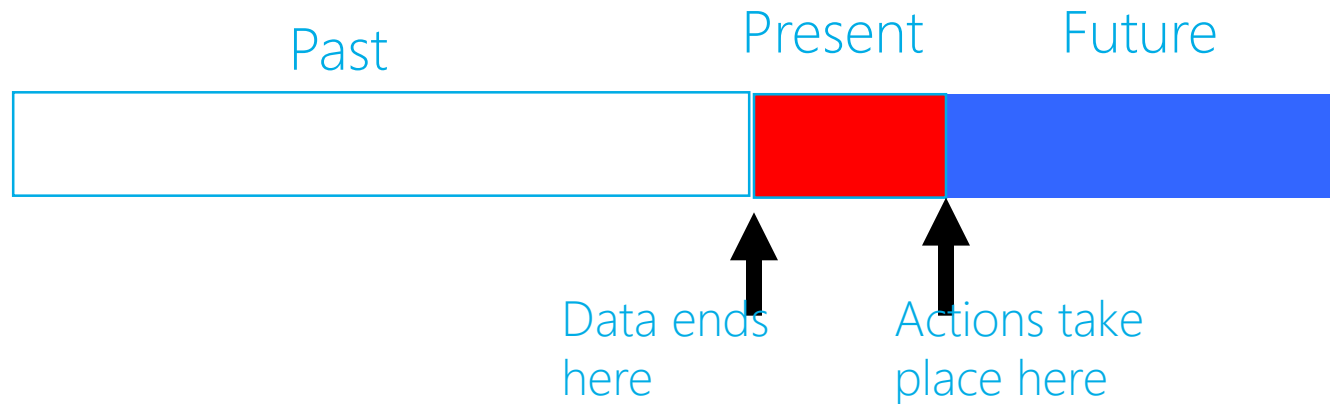


輸入變數(自變數)

輸出/目標變數(依變數)

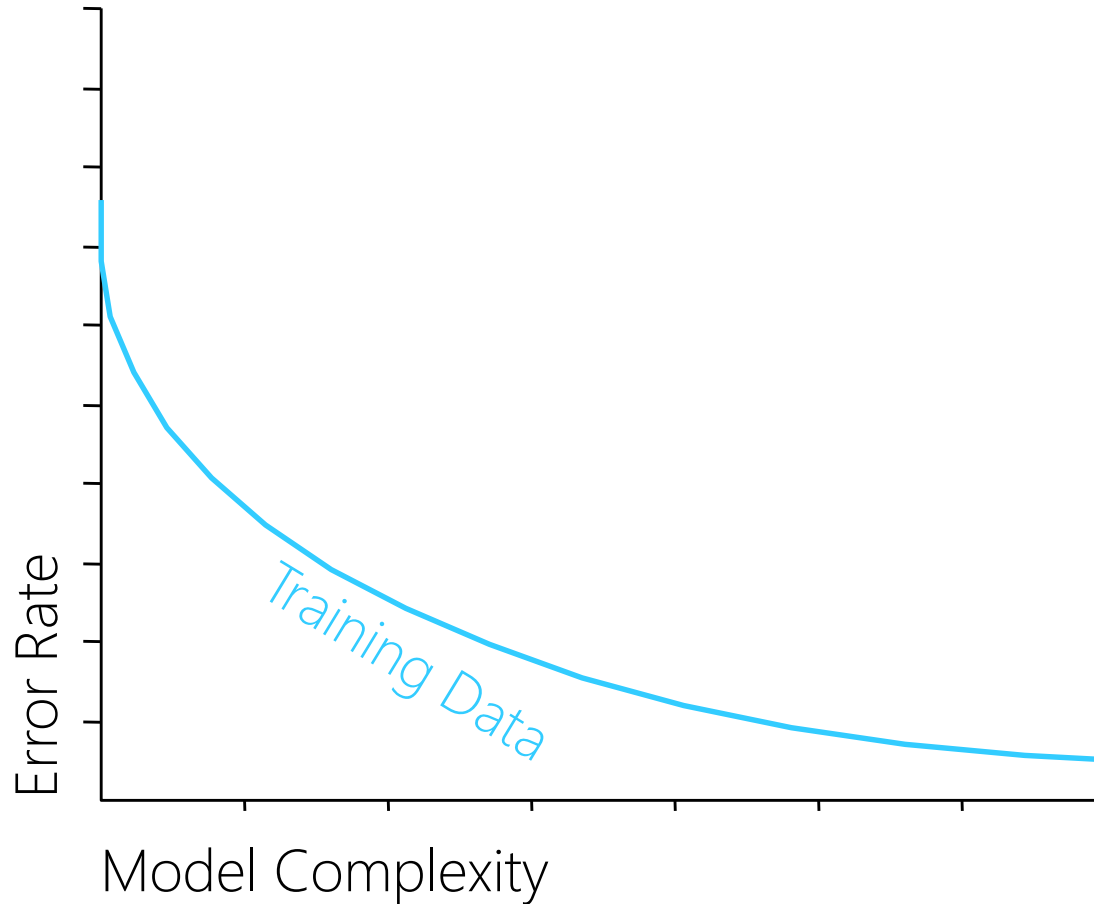
- 預測的前提：使用A預測B
 - A必須發生於B之前
 - 在B發生之前必須能取得A的資料

資料採礦是關於時間的

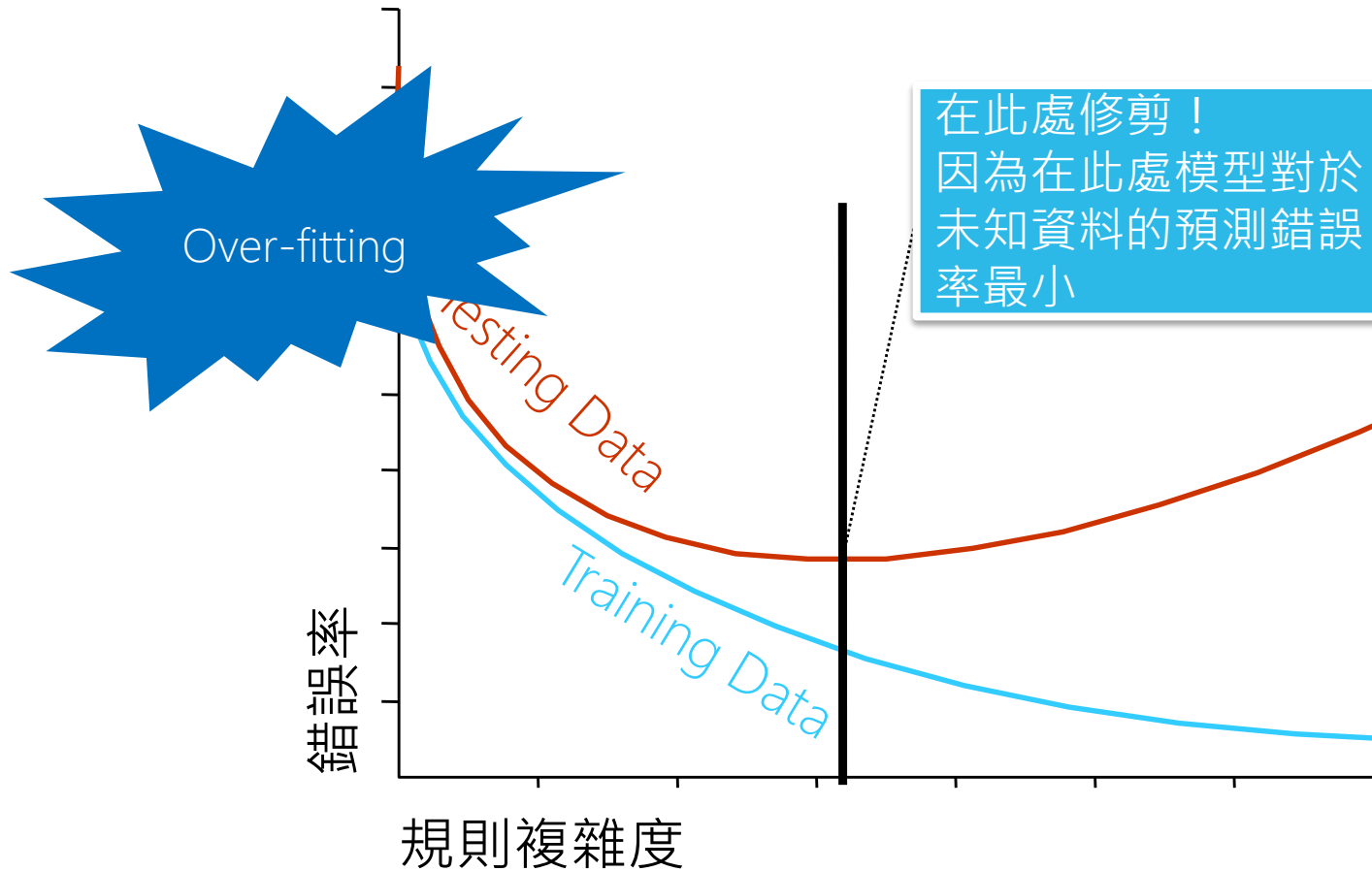


- 建立模型 **使用歷史資料**--結果是已知的
- 應用 (或是 評分) 模型 **使用現有的資料**
- 執行結果 **使用未來的資料**(下週或是下個月)--結果是未知的

模型穩定性挑戰



模型測試集的結果



樣本切割

Training

使用訓練組資料建立預測模型

Validation

使用鑑校組資料於建模階段評估模型

Test

使用測試組資料驗證模型

樣本切割



Properties

Split

Splitting mode

Split Rows

切割資料列

Fraction of rows in the first output dataset

0.6

切割比例

Randomized split

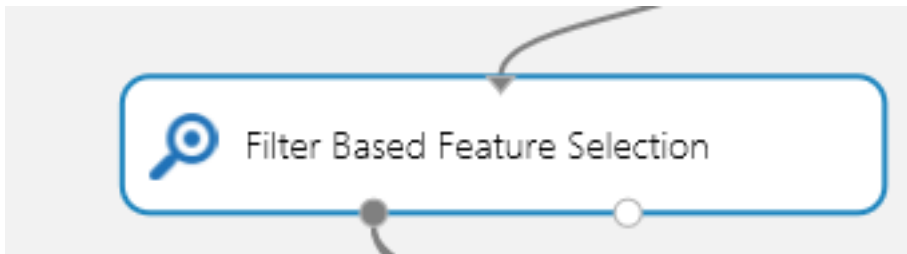
Random seed

0

Stratified split

False

變數選取



Properties

Filter Based Feature Selection

Feature scoring method

Mutual Information



變量選取方法

Operate on feature columns only



Target column

Selected columns:

Column names: Response

設定預測變數

Launch column selector

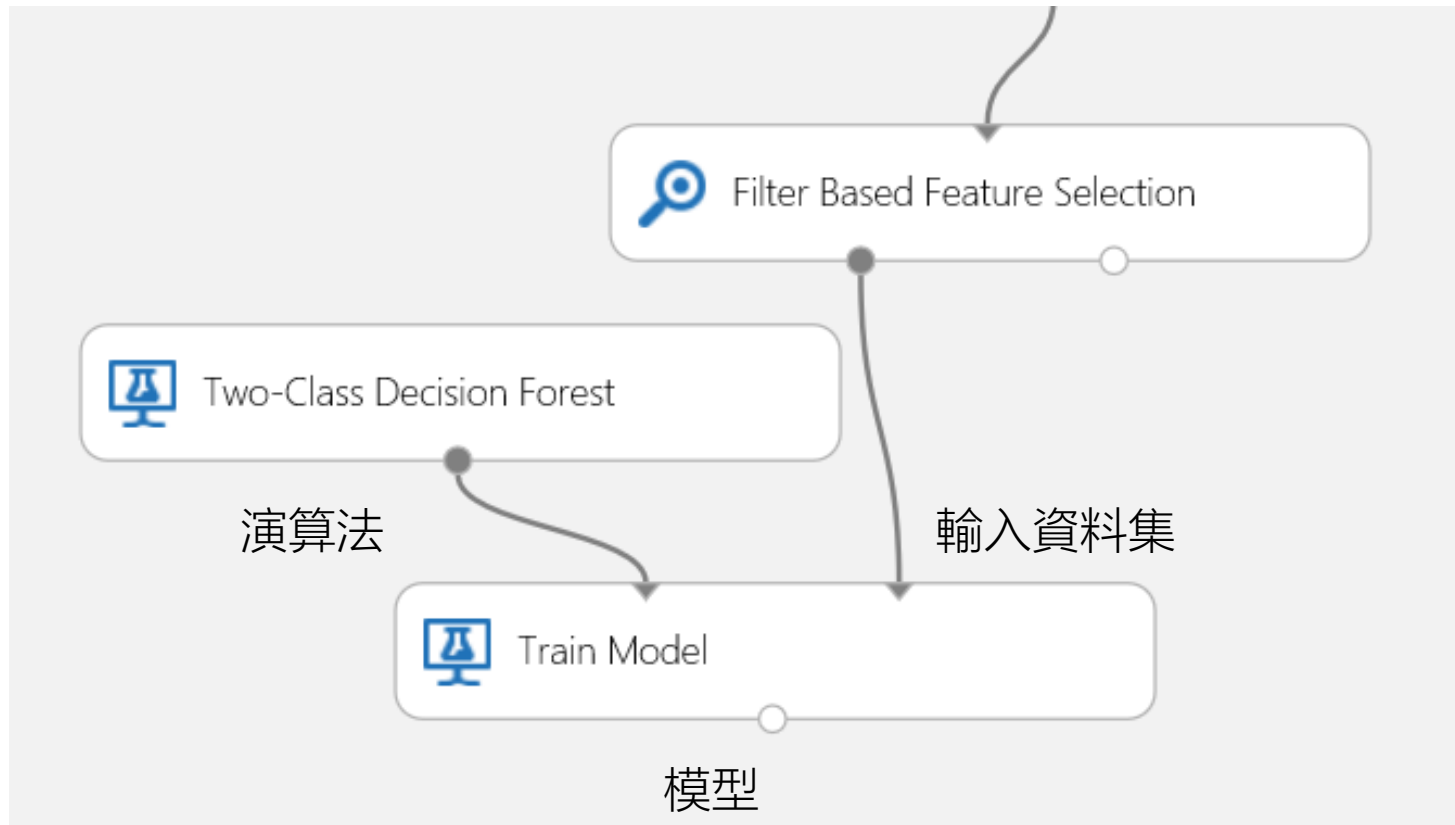
Number of desired features



10

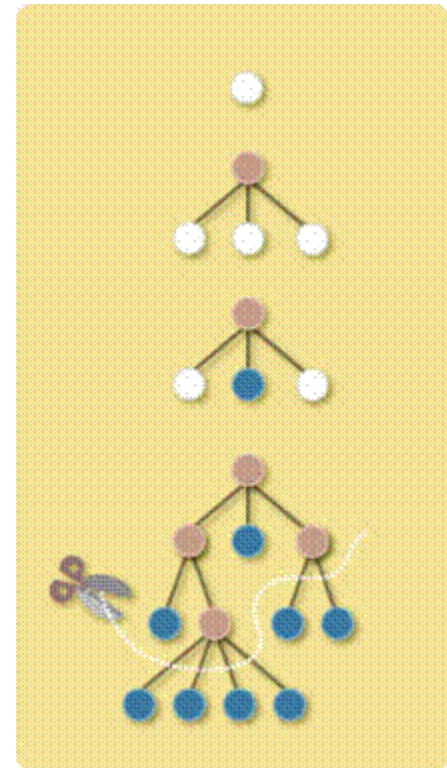
設定選入變數量

建立預測模型



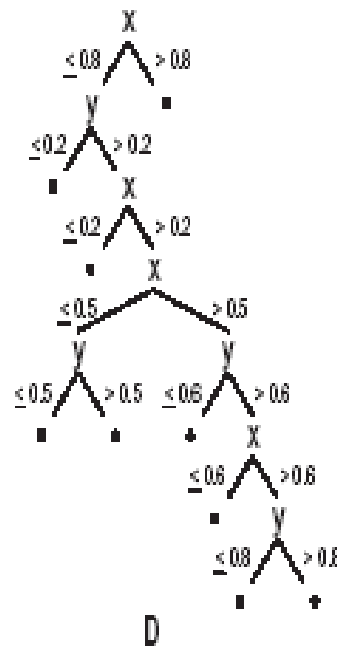
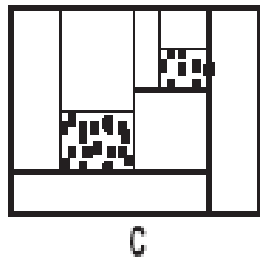
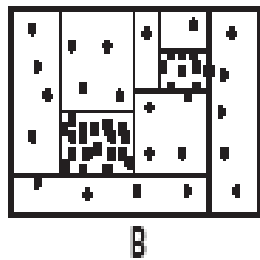
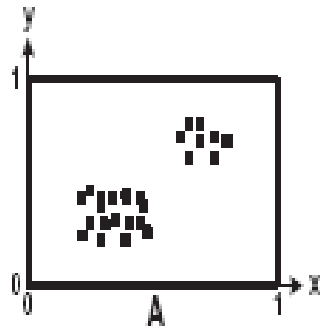
決策樹(Decision Tree)

- 透過樹狀規則呈現分類準則
- 找出最佳分岔點，使得資料的亂度最低
- 根據每個分岔的分布來決定預測機率

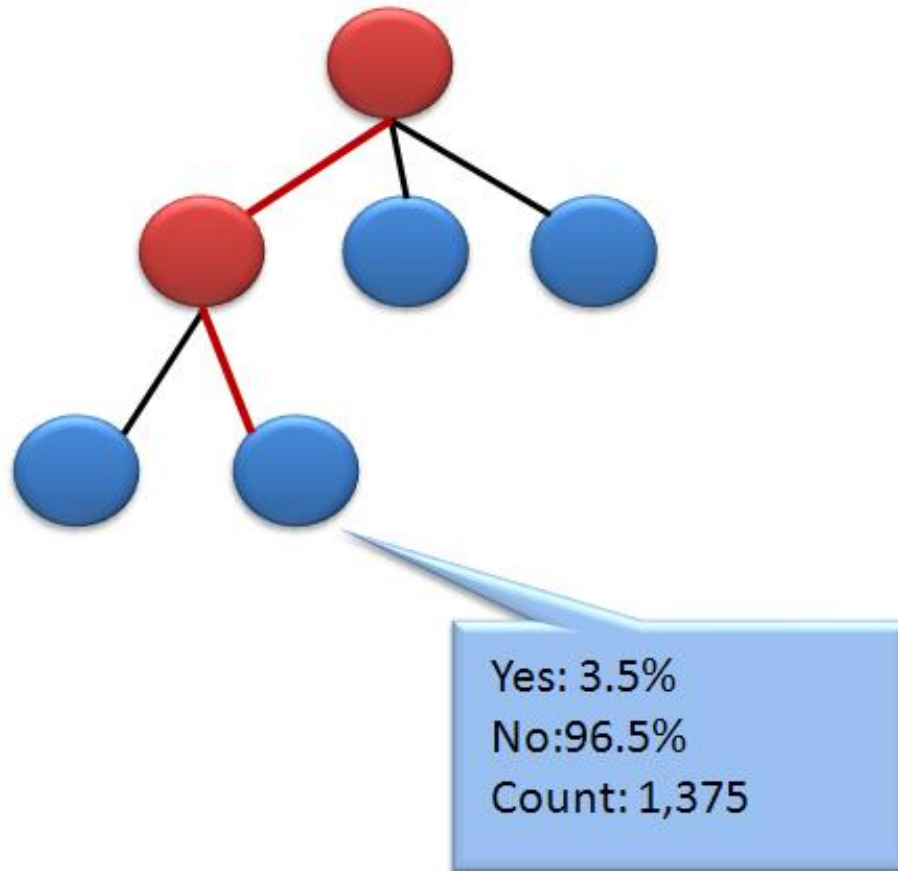


決策樹

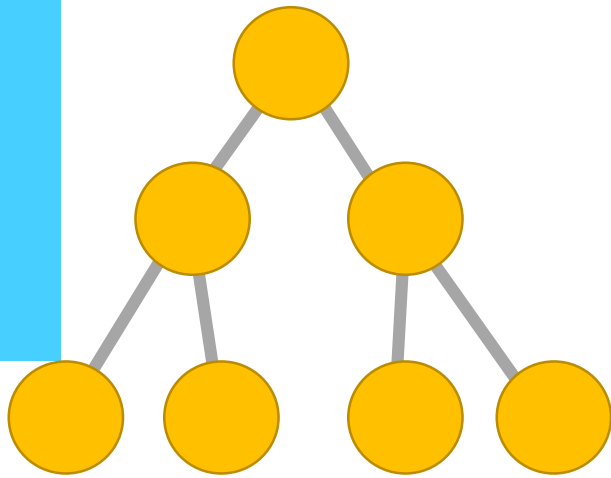
- 決策樹適合做為多維空間中方型切割以找尋最佳



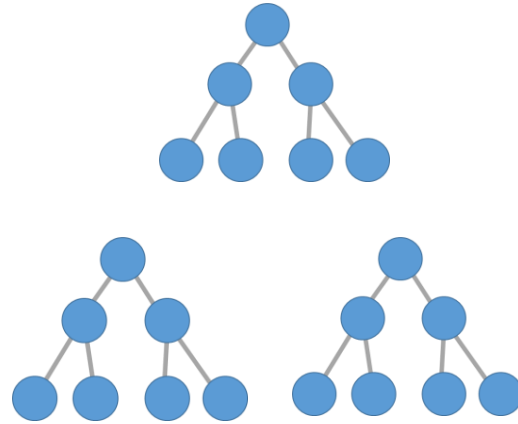
指派機率



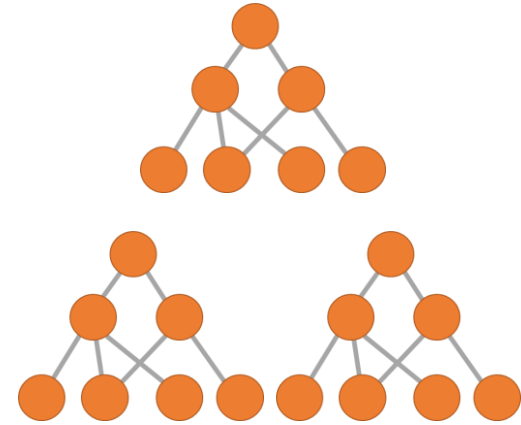
新型態的演算法



Decision Tree
決策樹



Decision Forest
決策森林

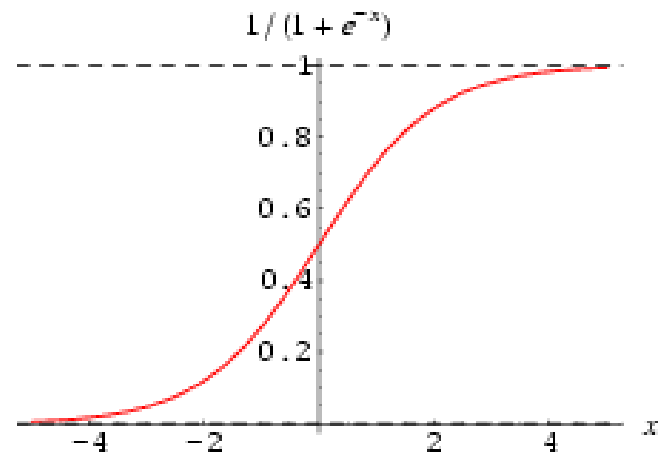


Decision Jungle
決策叢林

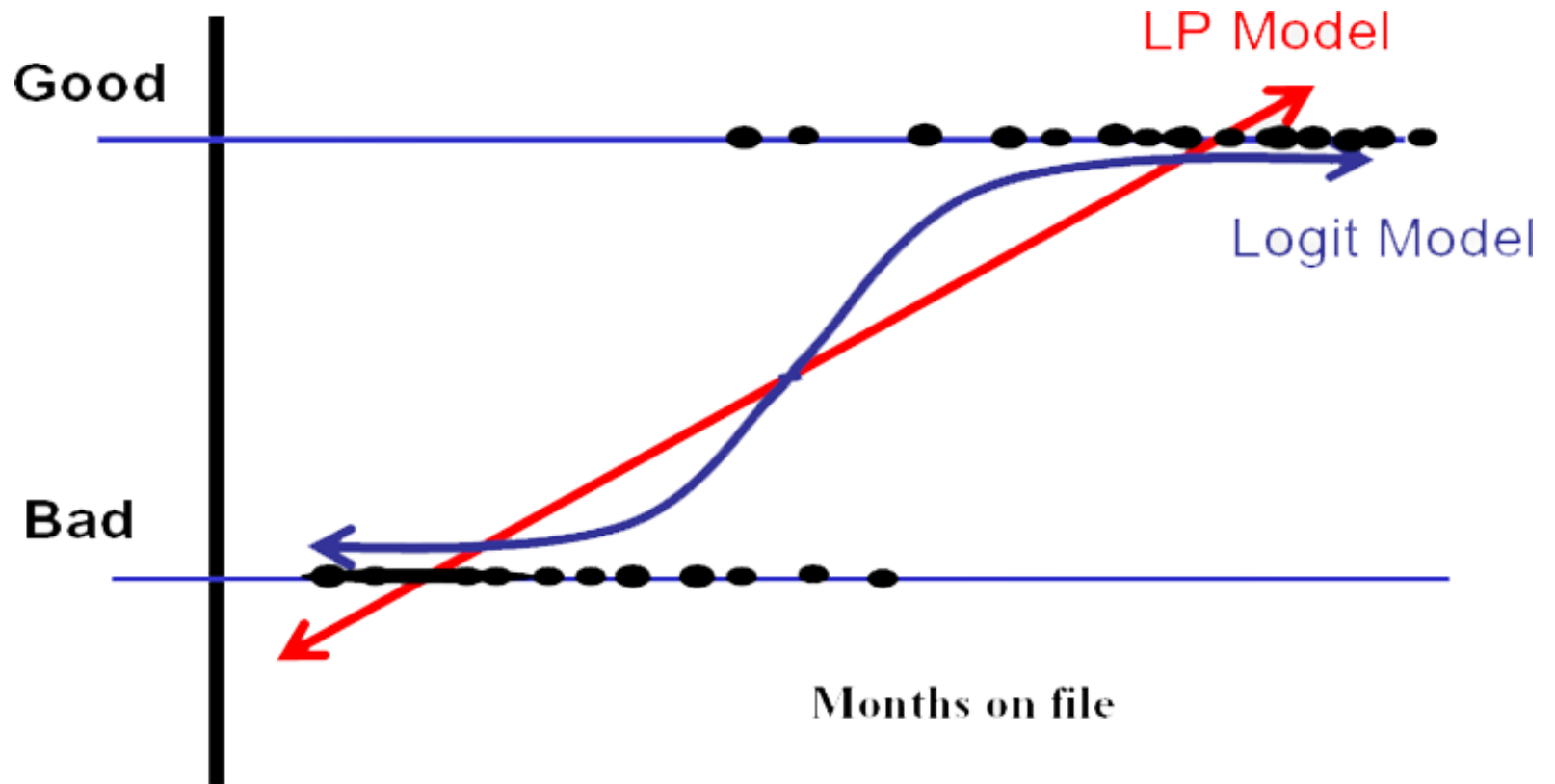
Logistic Regression

- $\ln(P/1-P)$ =線性方程式
- $P/(1-P)$ =odds rate
- 二元分類
- 輸入變數必須符合迴歸方程式，呈現單調遞增或是單調遞減趨勢

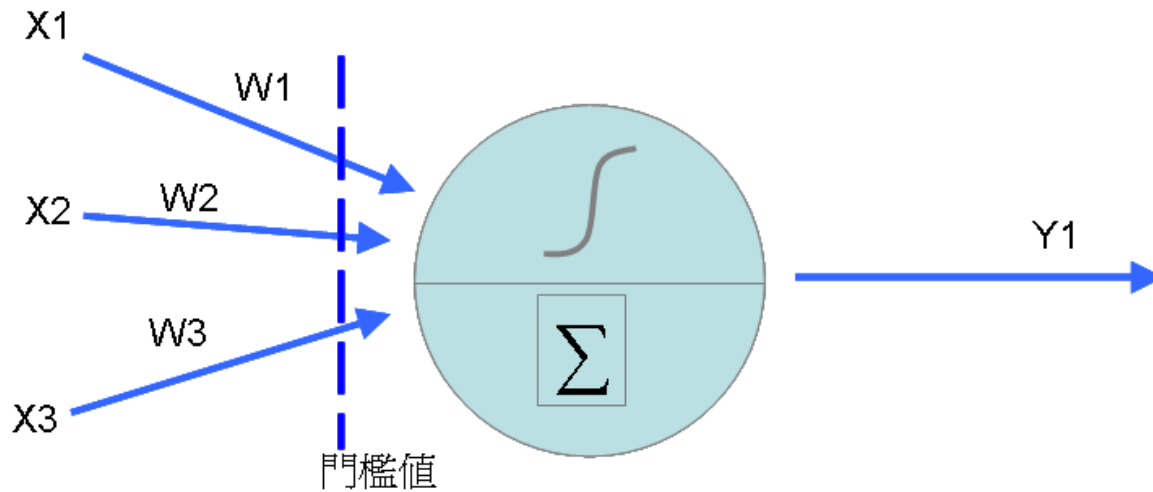
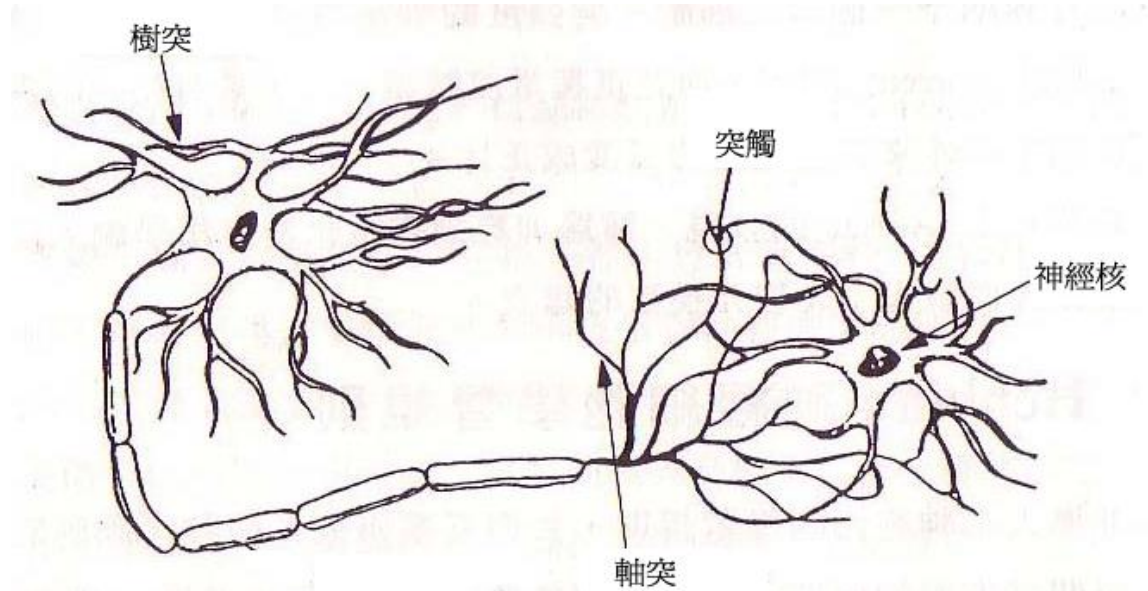
$$\ln \left[\frac{p}{1-p} \right] = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k$$



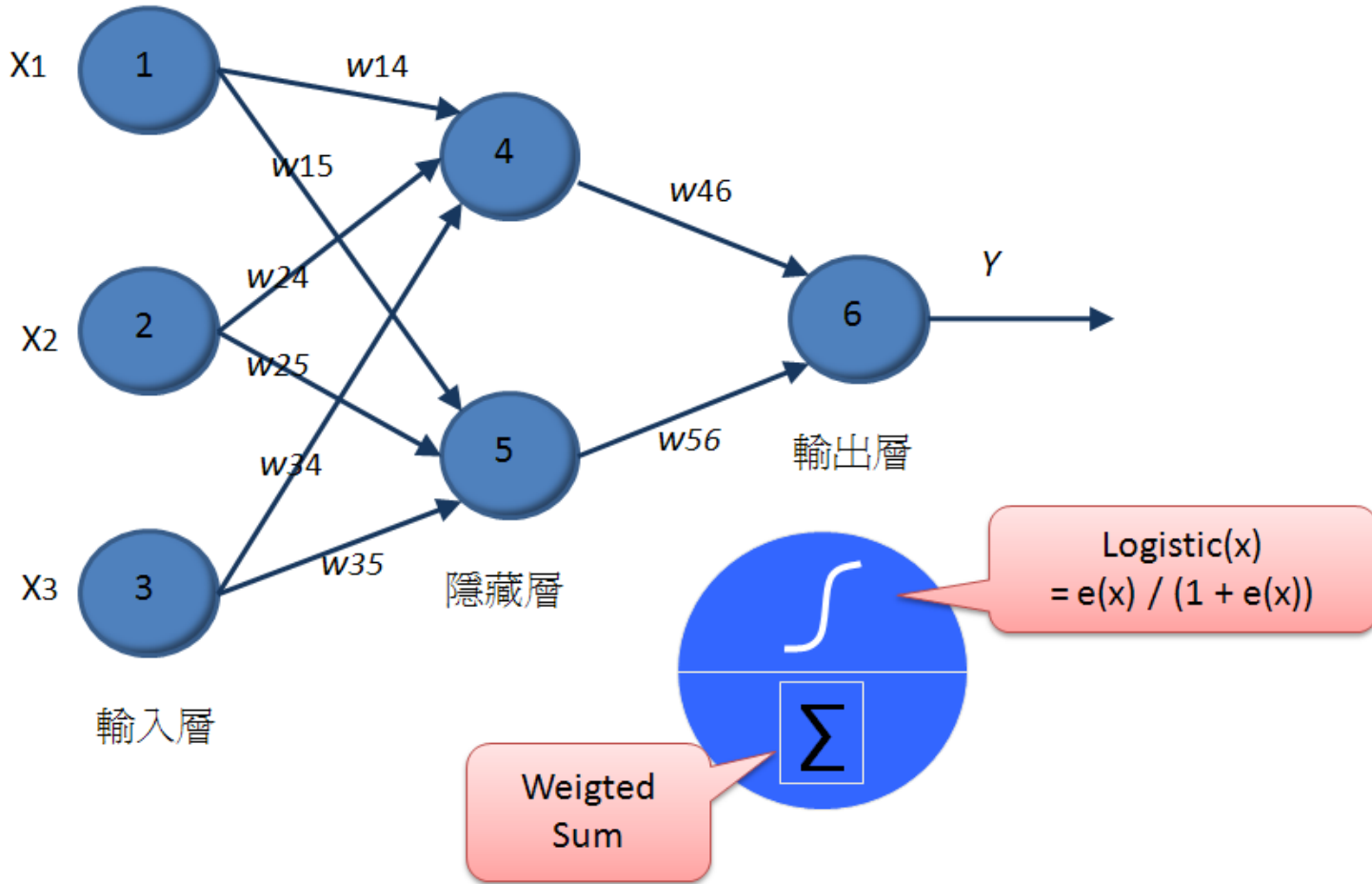
Logistic Regression



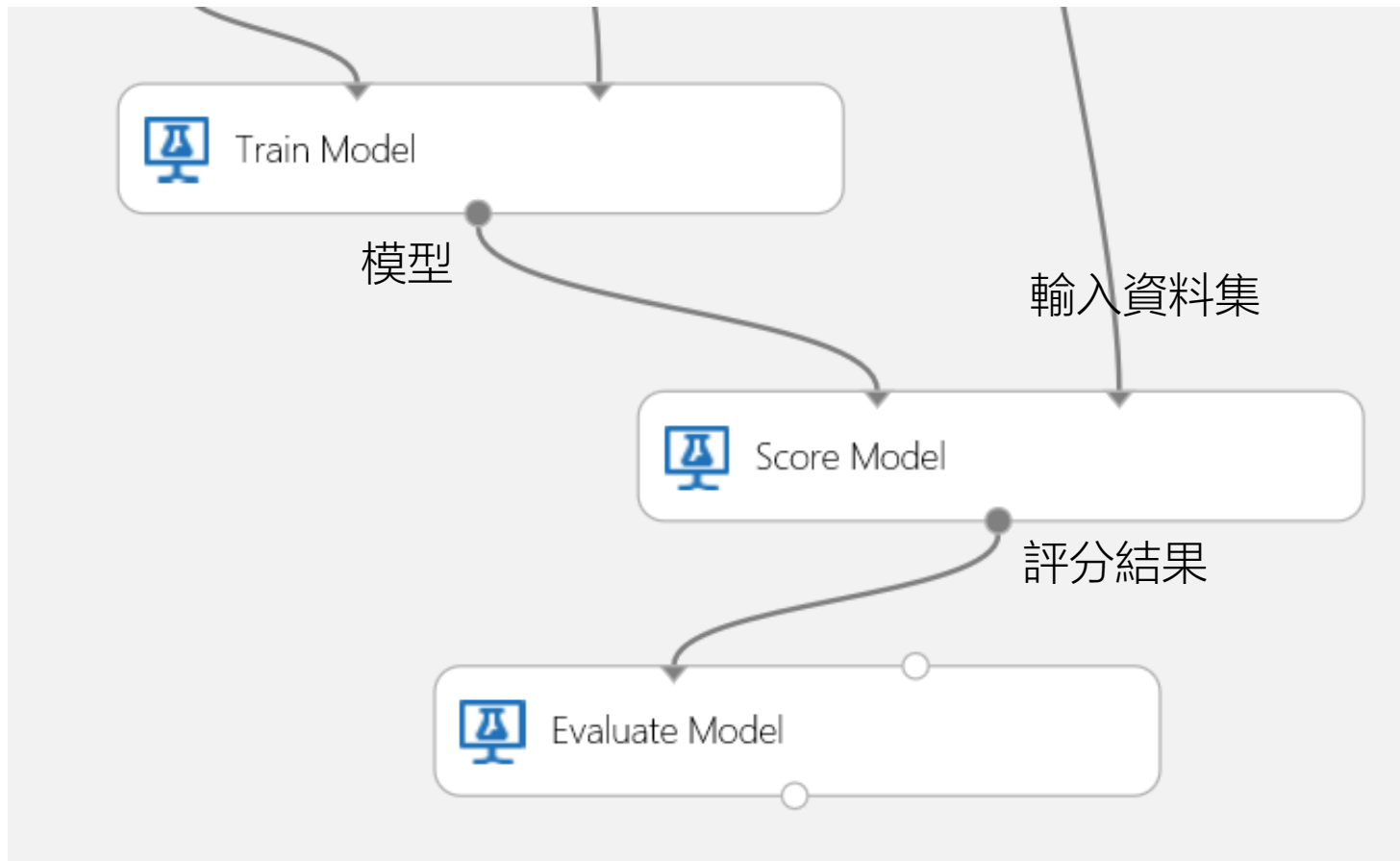
類神經網路



神經網路結構

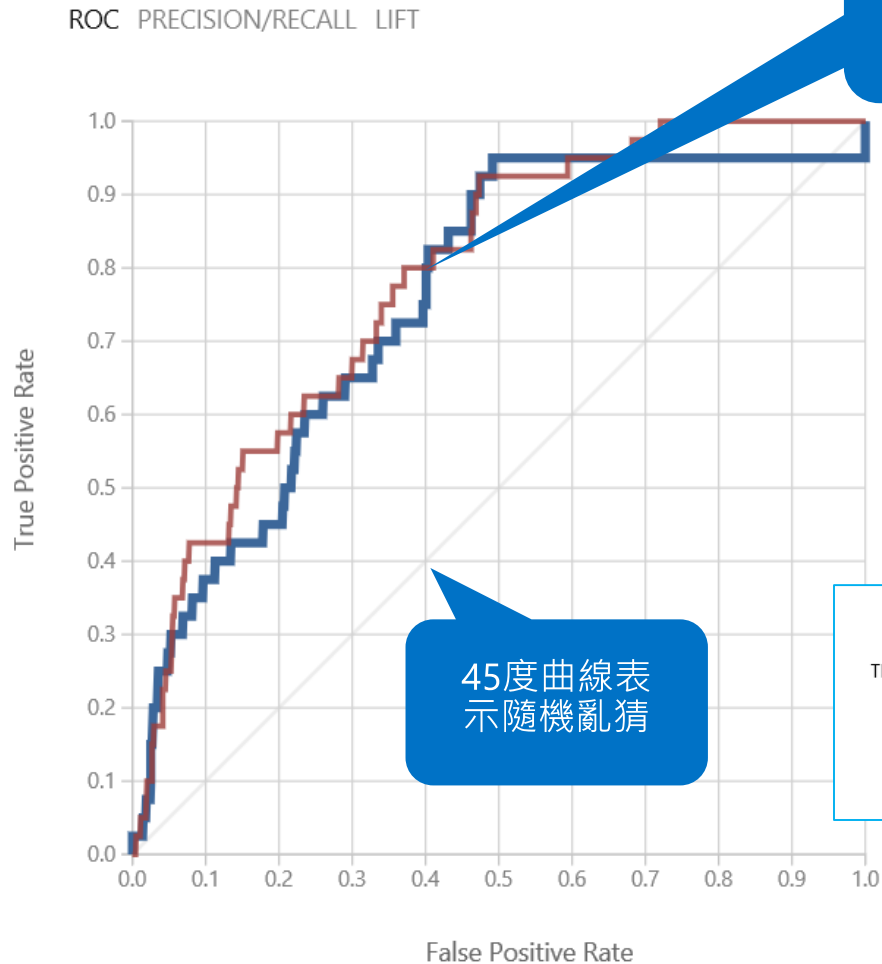


產生預測結果



評估預測模型

範例--客戶回應模型 > Evaluate Model > Evaluation results



透過資料採礦，
能讓前40%名單
捕捉到80%的
回應客戶

45度曲線表
示隨機亂猜

Threshold **0.5** Cumulative AUC **0.755**

AUC為曲線下面積，隨機為0.5，越大越接近完美模型

Prediction as a Service...

- 能將預測模型包裝為Web Services(單筆即時預測，多筆批次預測)
- 可透過C#, R以及Python等語言呼叫
- 將會是微軟物聯網策略中最重要的一環

The screenshot displays the Azure Machine Learning Studio interface for a web service named "範例--客戶回應模型". The interface includes a sidebar with navigation icons (Home, Experiment, Global, Settings) and a main content area with tabs for "DASHBOARD" and "CONFIGURATION". The "General" tab is active, showing the "Parent Experiment" and "範例--客戶回應模型". The "Description" field is empty, and the "API key" is displayed as a long alphanumeric string. Below the API key is a "Staging Services" table with columns for "URL", "TYPE", and "LAST U".

URL	TYPE	LAST U
API help page	REQUEST/RESPONSE	10/9/2014 1:15:26 AM
API help page	BATCH EXECUTION	10/9/2014 1:15:26 AM

The workflow diagram on the right shows a sequence of steps: "Train Model" (checked), "Score Model" (checked), and "Project Columns" (checked). The "Score Model" step is highlighted with a green checkmark and a refresh icon, indicating it is the current step being viewed.

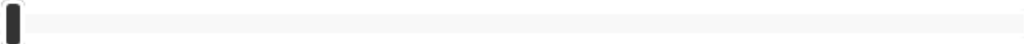
Azure Machine Learning 計費模式

機器學習

ML Studio

ML API 服務

只有執行實驗或模型訓練之作用中的 Studio 服務使用會列入計費。如需預估使用時數，請參閱[常見問題集](#)。

	1小時	NT\$11.80
實驗小時		

機器學習

ML Studio

ML API 服務

同時套用預測和預測小時的度量。預測小時的計算方式是預測數量乘以每次預測的時間。如需預估預測時間的資訊，請參閱[常見問題集](#)。

	1百萬	NT\$5,584.95
0-100 K 100-1000 K 1 百萬到 1 億 1 億到 10 億 預測		
	0s	NT\$0.00
0-10 10-60 60-600 600-3600 每次預測的秒數		

A woman with her hair in a ponytail, wearing a black and white striped shirt, is pointing at a tablet. A man with glasses and a suit is looking at the tablet. The scene is overlaid with a semi-transparent orange rectangle containing the text '客戶區隔模型'.

客戶區隔 模型

傾向模型與相似模型?

- 傾向模型(Propensity Model)
 - 輸入->輸出
- 相似模型(Look-like Model)
 - 沒有輸出變數
 - 找出相似結構...

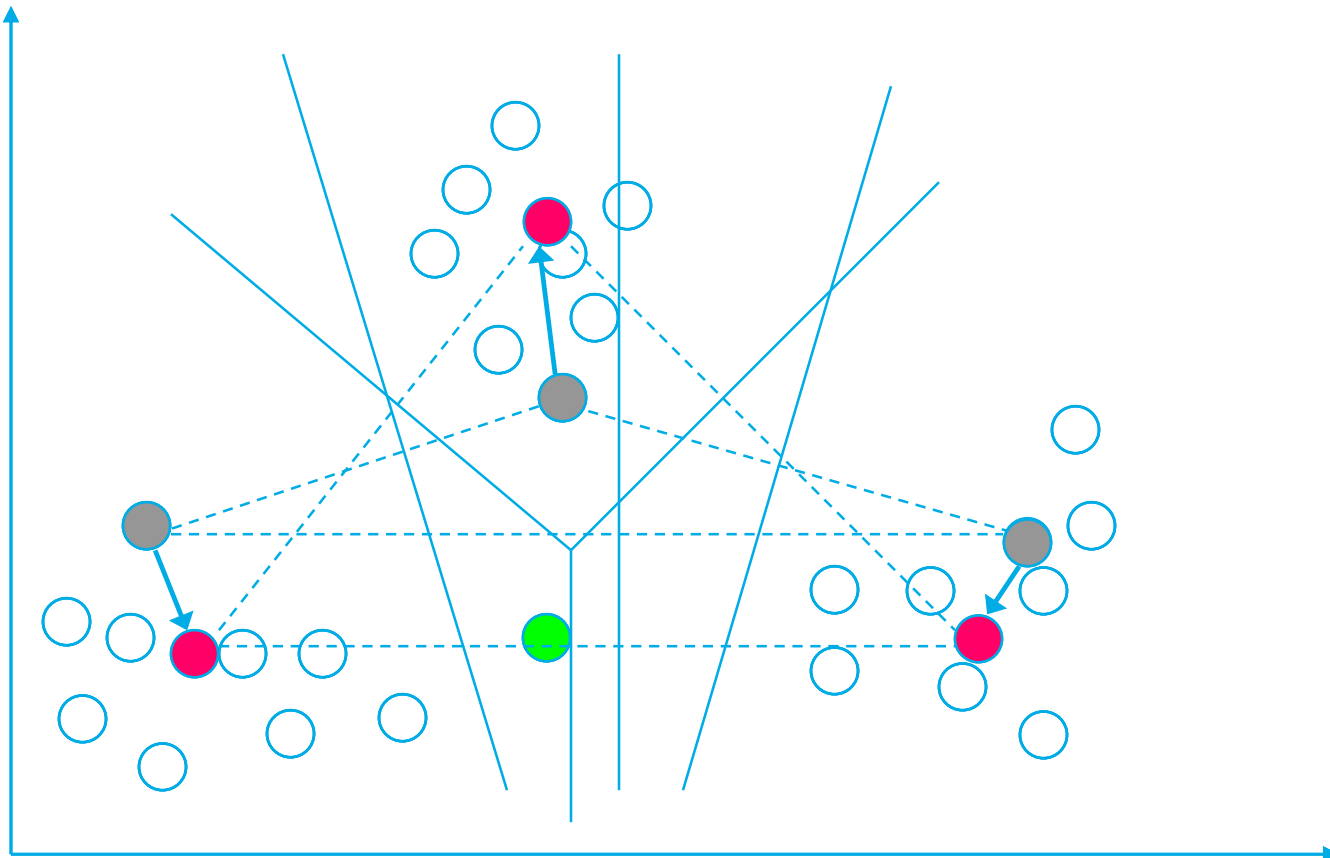
Look-like Model



K-means演算法

- 最常使用的分群演算法
- K =分群數
- 透過隨機指派質量中心，透過不斷重新指派成員、計算新中心，來趨近於分群邊界

K-means



建立分群模型



A photograph of a woman with brown hair in a ponytail, wearing a black and white striped shirt, pointing at a tablet. A man with grey hair and glasses, wearing a dark suit and a light blue shirt, is looking at the tablet. The scene is set in an office environment. An orange semi-transparent rectangle is overlaid on the image, containing the text '整合R語言'.

整合R語言

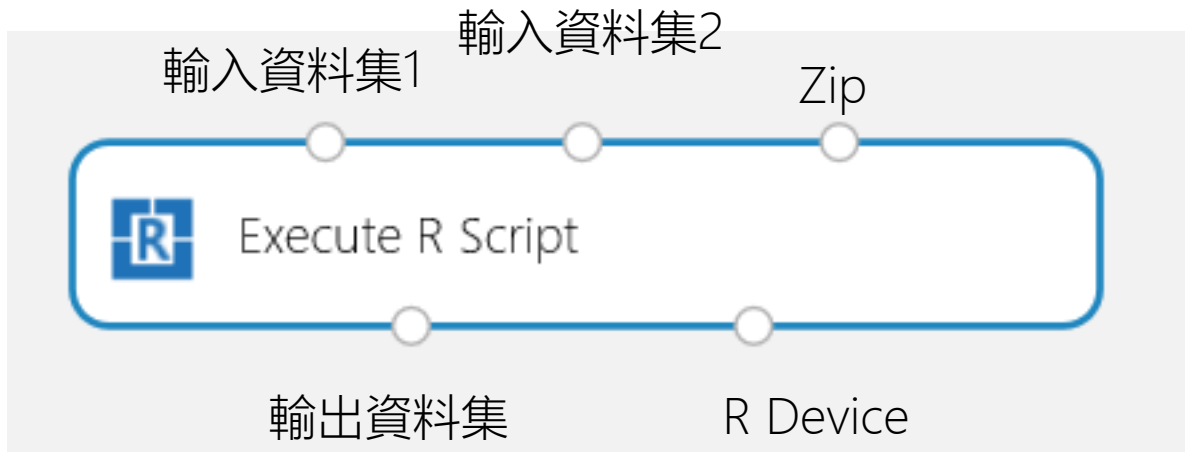
R語言

- R語言是由紐西蘭奧克蘭大學的Ross Ihaka和Robert Gentleman開發的開源統計語言
- 能夠進行統計、預測模型以及資料視覺化

<http://www.r-project.org/>



如何整合R語言



```
dataset1 <- maml.mapInputPort(1)  
dataset2 <- maml.mapInputPort(2)
```

輸入

```
data.set = rbind(dataset1, dataset2);
```

```
maml.mapOutputPort("data.set");
```

輸出

檢視預裝的R Package

```
data.set<-data.frame(installed.packages())  
maml.mapOutputPort("data.set");
```

Microsoft Azure Machine Learning | Enter feedback here | AsiaMiner | Menu

檢查Azure Machine Learning預裝的R Packages數量

Finished running ✓

檢查Azure Machine Learning預裝的R Package... > Execute R Script > Result Dataset

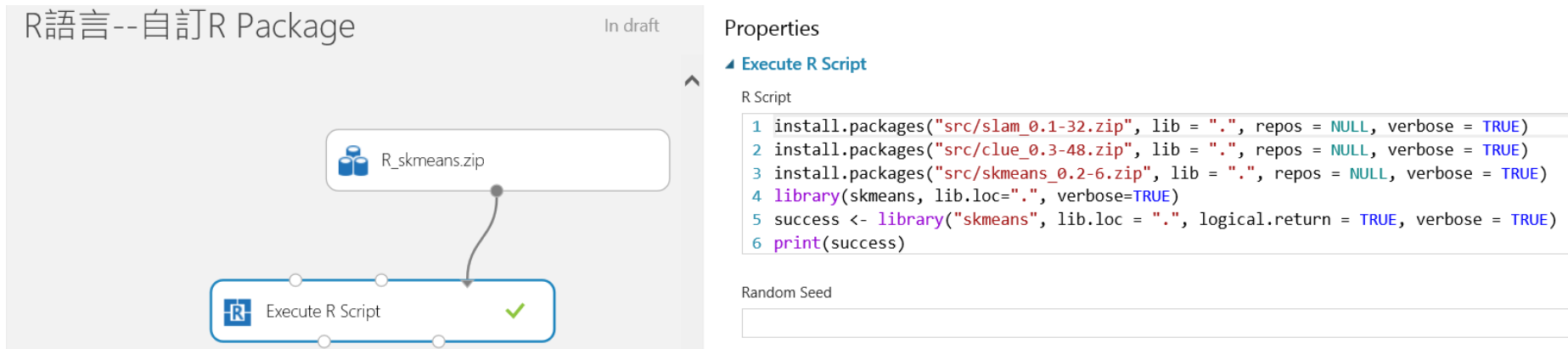
410 rows | 16 columns

view as [Bar Chart]

	Package	LibPath	Version	Priority	Depends	Imports	LinkingTo	Suggests	Enhances
Mean									
Median									
Min									
Max									
Standard Deviation									
Unique Values	410	1	317	3	248	128	9	188	
Missing Values	0	0	0	381	77	255	386	200	3
Feature Type	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical
	abc	C:/ThirdParty/library	1.8		R (>= 2.10), nnet, quantreg, MASS				
	abind	C:/ThirdParty/library	1.4-0		R (>= 1.5.0)				
	actuar	C:/ThirdParty/library	1.1-6		R (>= 2.6.0)	stats, graphics		MASS	

如何加入自訂R Packages

- 在電腦本機安裝想要的R Package
- 將R Package原始的Zip檔壓縮為Zip
- 上傳至Azure Machine Learning



R語言--自訂R Package In draft

R_skmeans.zip

Execute R Script ✓

Properties

Execute R Script

R Script

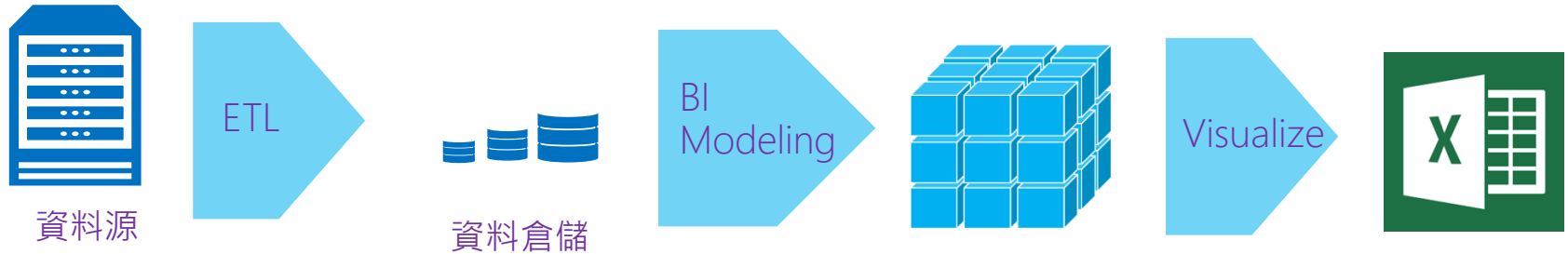
```
1 install.packages("src/slam_0.1-32.zip", lib = ".", repos = NULL, verbose = TRUE)
2 install.packages("src/clue_0.3-48.zip", lib = ".", repos = NULL, verbose = TRUE)
3 install.packages("src/skmeans_0.2-6.zip", lib = ".", repos = NULL, verbose = TRUE)
4 library(skmeans, lib.loc=".", verbose=TRUE)
5 success <- library("skmeans", lib.loc = ".", logical.return = TRUE, verbose = TRUE)
6 print(success)
```

Random Seed

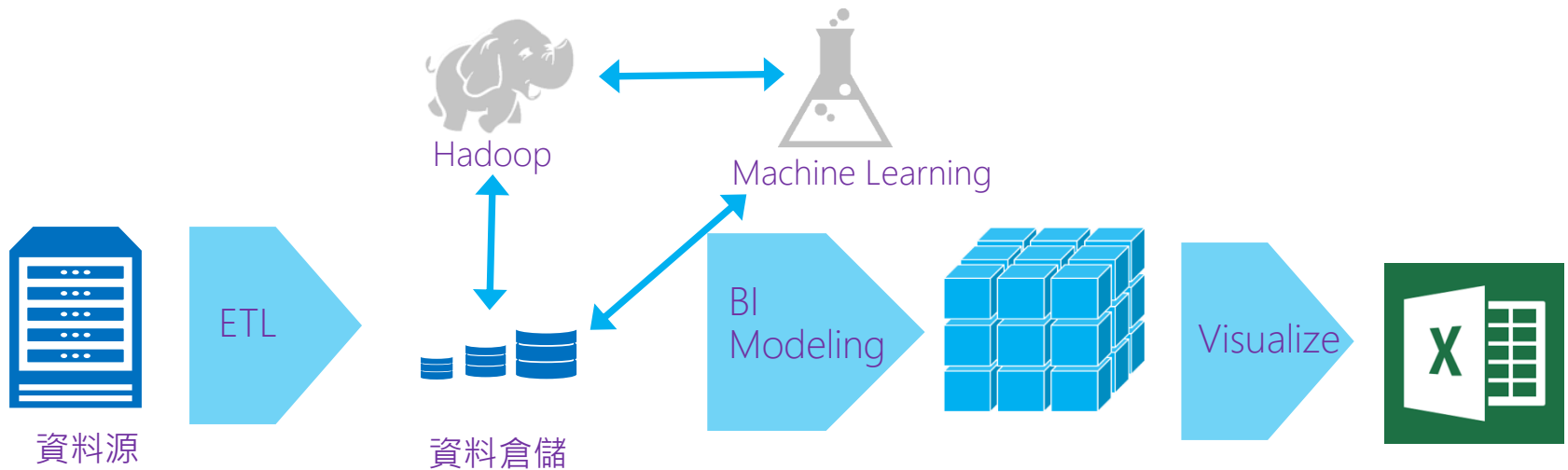
A woman with her hair in a ponytail, wearing a black and white striped shirt, is pointing at a tablet. A man with glasses and a suit is looking at the tablet. The scene is overlaid with a semi-transparent orange rectangle containing the title text.

整合Power BI for Office 365

Traditional BI



Modern Predictive BI

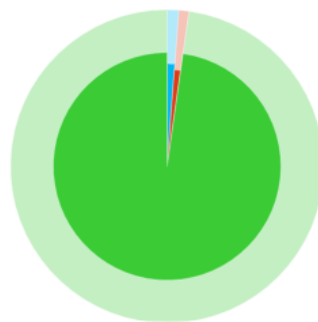
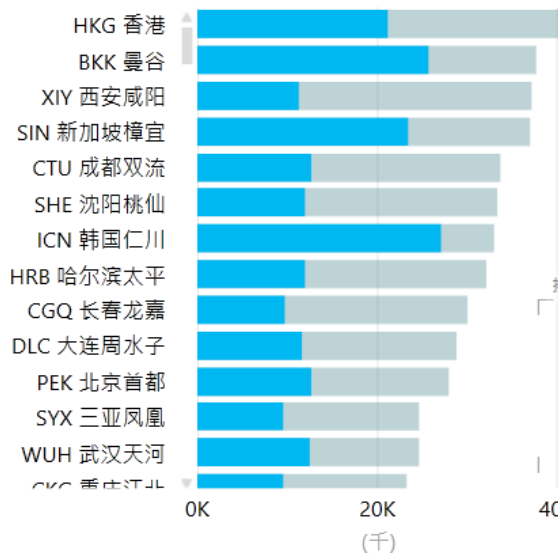


整合Power BI

客户细分搭机行为

客户细分

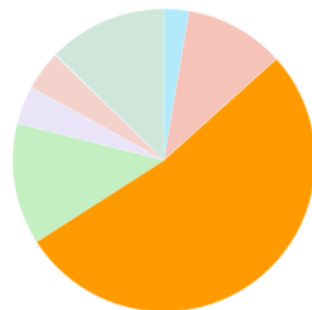
- 细分1
- 细分2
- 细分3
- 细分4
- 细分5
- 细分6
- 细分7
- 细分8
- 细分9
- 细分10
- 细分11
- 细分12



舱位

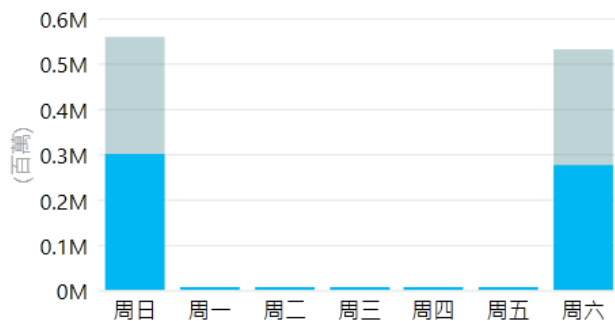
- C
- F
- W
- Y

排序依据 舱位类型 递增



舱位类型

- 中转舱
- 正常公布价
- 折扣舱
- 里程兑换舱
- 超经舱位
- 团队舱
- 职工免票舱
- 预售舱

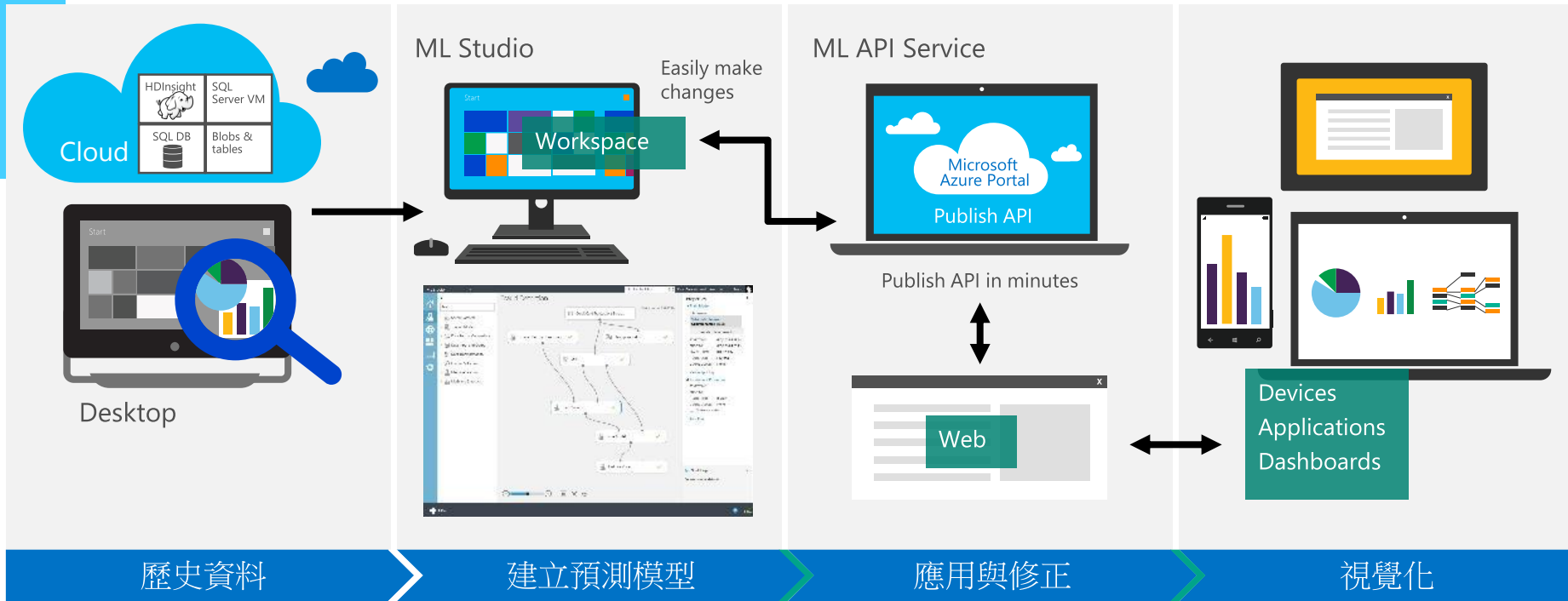


出发时段

- 中午
- 半夜
- 早上
- 晚上



完整的數據分析生命週期



Q&A

