



bigData
Spark
Forum

競賽經驗談：從國內 競賽到 Kaggle 實戰

Yung-Chuan Lee
2017.03



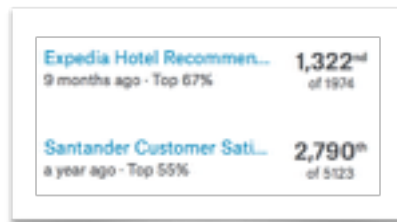
我的養成

104 上半年度產投課程 - Big Data 資料分析概論班 2015-08-30(日) 至 2015-09-13(日)

行雲流水軟體 AppCloudom Soft
Spark/Scala大數據分析技術與應用
2015年12月21日



- 2015/06 決定投入大數據領域
- 2015/09 參加胡老師「104上半年度產投課程 – Big Data 資料分析概論班」
- 2015/12 參加胡老師「Spark/Scala大數據分析技術與應用」
- 2016/03 開始在Kaggle修練經驗值
- 2016/08 參加「電子商務巨量資料分析競賽」 – 季軍
- 2017/02 參加論譚課程「SparkR 大數據分析實務」



關於資料競賽

■ 資料分析競賽

- 主辦單位提供去識別化資料、預測標的、競賽平台
- 參賽者在平台上透過資料分析方法，最佳化預測能力

■ 資料應用競賽

- 使用現有資料進行加值應用

Etu
HADOOP
Competition
2016+udn

2016 HackNTU
To think, to hack.

競賽過程 = 資料分析流程 + 成果簡報

EHC 2016 決賽結果

學生組

名次	隊名	總分	準確率	程式效能	平行運算	分析作法	創新作法
1	我不知道	44.26	4.60	5.46	15	9.60	9.60
2	CSIE	43.34	6.30	7.24	15	7.60	7.20
3	Watchowl	38.28	3.70	13.58	15	2.80	3.20
4	B!hance	33.50	5.70	15	0	6.80	6.00
5	King	32.80	4.00	5.00	15	4.00	4.80

社會組

名次	隊名	總分	準確率	程式效能	平行運算	分析作法	創新作法
1	大GA	48.84	3.70	12.54	15	8.80	8.80
2	數據好好玩	48.66	7.10	14.56	15	6.00	6.00
3	李氏兄弟	46.97	5.40	14.97	15	6.00	5.60
4	DDA	43.00	5.00	5.00	15	9.20	8.80
5	mick	40.40	5.60	15.00	15	2.80	2.00

資料分析流程

• 資料清理

資料前置處理
(Data Pre-processing)
需要“領域知識”

1

原始資料
Raw Data

傳統資料分析 (R)
vs.
Big Data 資料分析 (SparkR)

• 探索式資料分析(EDA)

2

資料視覺化
(Data Visualization)

特徵向量擷取
(Feature Vector's
Determination)

傳統資料分析流程：R

資料分析方法：

- 機率模型
- 統計模型
- 資料探勘 (Data Mining)

3

機器學習訓練與測試流程
(Machine Learning Pipeline for Training & Testing)

Big Data 資料分析流程：SparkR

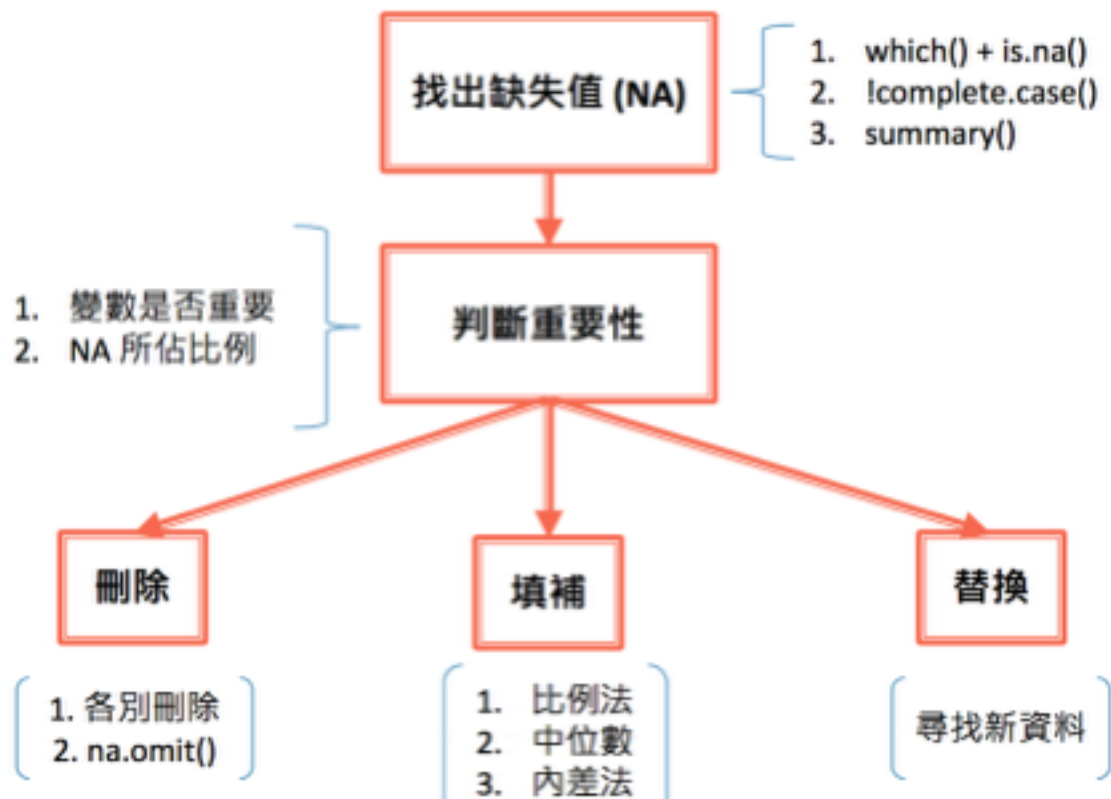
調校(Tuning)

- Feature Engineering
- 演算法參數調整
- 換演算法

模型校能評估 (Evaluation)

- AUC
- F-measure
- RMSE

缺失值 (NA) 處理

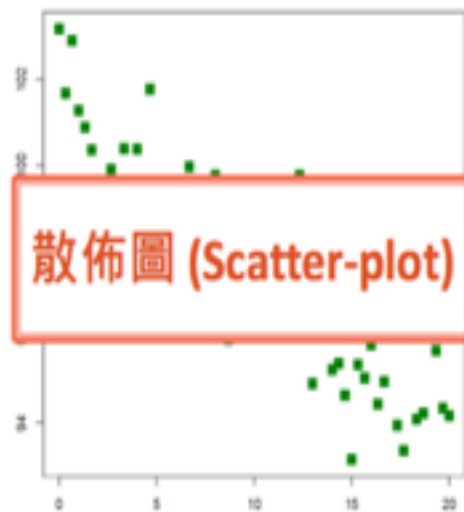
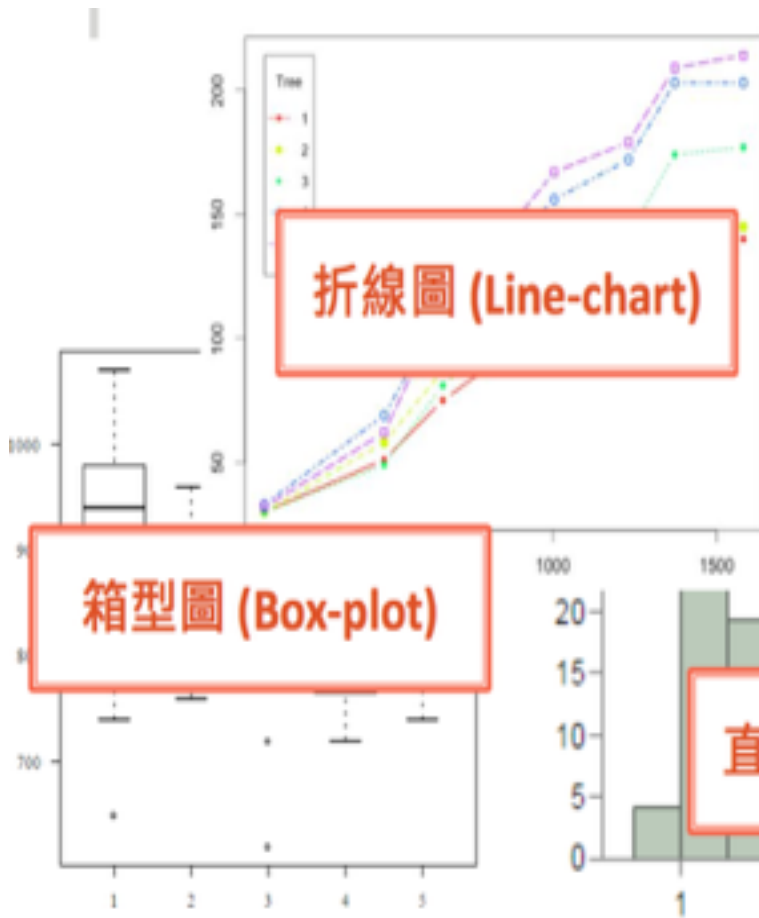


探索式資料分析(Exploratory Data Analysis)

■ 什麼是 EDA ?

- 一種初步分析的方法 (或態度), 主要是 透過畫圖的方式，達到三個主要的目的
 - 最大化對資料的了解
 - 找出重要的變數
 - 發現 outliers 或異常數值
- 不做過度假設地從原始數據看出隱含意義

EDA 常用的 視覺化方式



特徵向量擷取(Feature Extraction)

■ Feature Extraction的目的

- 經過資料清理及 EDA 後，選取分析所需之 $Feature$
 - 一般不會把 ID 、序號等拿來分析
- 依演算法特性處理後作為演算法的 $input$
 - EX : 線性演算法前將 $Feature$ 作 $normalized$
 - EX : $Spark.naiveBayes$ 只接受 $categorical\ feature$

模型訓練 (Model Training)

- 擷取後的Feature輸入選定的演算法產生Model

Spark MLlib		
	Categorical Qualitative	Continuous Quantitative
Unsupervised Extracting structure	Clustering K-means	Dimension Reduction Singular Value Decomposition (SVD) Principal Component Analysis (PCA)
Supervised Making prediction	Classification Naive Bayes Decision Trees Ensembles of Trees (Random Forests and Gradient-Boosted Trees)	Regression linear models Support Vector Machines logistic regression linear regression
Recommender Associating user item	Collaborative Filtering Alternating Least Squares (ALS)	
Optimization Finding minima		Optimization Stochastic Gradient Descent Limited-memory BFGS (L-BFGS)
Feature Extraction Processing text	Feature Extraction Transformation TF-IDF - Word2Vec Standard Scaler - Normalizer	

模型校能評估(Evaluation)

- 透過特定方法評估**Model**的預測效果
 - *Clustering* : 組內最小平方和(WCSS)
 - *Classification* : *AUC*、*F-Measure*、*Accuracy*
 - *Regression* : *RMSE*(*Root Mean Square Error*)

分類問題常用校能評估指標

		predicted condition			
total population		prediction positive	prediction negative	Prevalence $= \frac{\Sigma \text{condition positive}}{\Sigma \text{total population}}$	
true condition	condition positive	True Positive (TP)	False Negative (FN) (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{TP}}{\Sigma \text{condition positive}}$	False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{FN}}{\Sigma \text{condition positive}}$
	condition negative	False Positive (FP) (Type I error)	True Negative (TN)	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{FP}}{\Sigma \text{condition negative}}$	True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{TN}}{\Sigma \text{condition negative}}$
$\text{Accuracy} = \frac{\Sigma \text{TP} + \Sigma \text{TN}}{\Sigma \text{total population}}$		Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{TP}}{\Sigma \text{prediction positive}}$	False Omission Rate (FOR) $= \frac{\Sigma \text{FN}}{\Sigma \text{prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) $= \frac{\text{LR}^+}{\text{LR}^-}$
		False Discovery Rate (FDR) $= \frac{\Sigma \text{FP}}{\Sigma \text{prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\Sigma \text{TN}}{\Sigma \text{prediction negative}}$	Negative Likelihood Ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	

- ▶ Recall(實際為1，也被正確判定為1的比例) $\Rightarrow \text{TPR} = \text{TP} / (\text{TP} + \text{FN})$
- ▶ Precision(判定為1，實際也為1的比例) $\Rightarrow \text{PPV} = \text{TP} / (\text{TP} + \text{FP})$
- Accuracy(正確答對的比例) $\Rightarrow \text{ACC} = (\text{TP} + \text{TN}) / (\text{P} + \text{N})$

模型調校(Model Tuning)

- 演算法參數調校

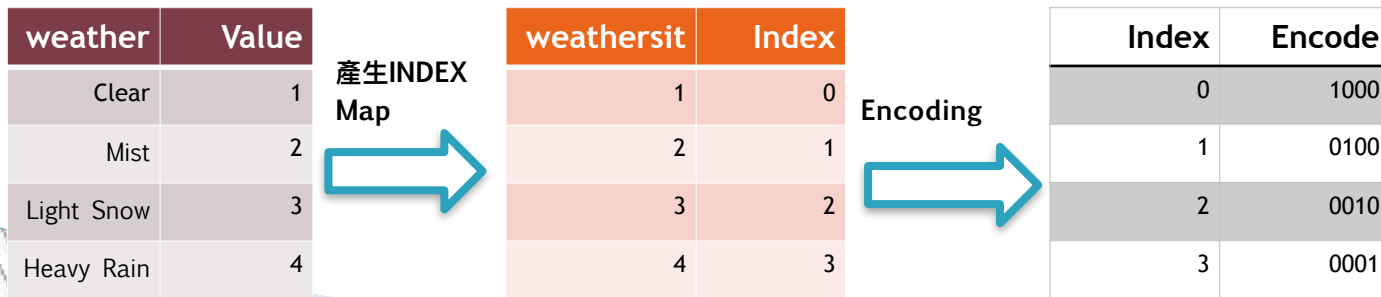
- *EX*：嘗試*RandomForest*的*Tree*數量、樹深度及分枝數

- 特徵工程(**Feature Engineering**)

- 取得更多特徵值、改變特徵值的呈現(*EX:One-hot encoding*)來提升演算法效能

Categorical Features的處理










- 部份演算法(如LogisticRegression)對Categorical欄位處理能力較差，因此針對類別型態的Feature需作one-of-k(one-hot) encoding
- One-of-K encoding:
 - 長度為N的整數陣列(N=欄位類別的數量)
 - 陣列中類別對應的index設為1，其它為0



Let's Kaggle!!!

■ <https://www.kaggle.com/competitions>

kaggle

	Intel & MobileODT Cervical Cancer Screening Which cancer treatment will be most effective? <i>Featured</i> · 3 months to go	\$100,000 166 teams
	Google Cloud & YouTube-8M Video Understanding Challenge Can you produce the best video tag predictions? <i>Featured</i> · 2 months to go	\$100,000 340 teams
	Quora Question Pairs Can you identify question pairs that have the same intent? <i>Featured</i> · 2 months to go	\$25,000 721 teams
	Two Sigma Connect: Rental Listing Inquiries How much interest will a new rental listing on RentHop receive? <i>Recruitment</i> · a month to go	Jobs 1,540 teams
	March Machine Learning Mania 2017 Predict the 2017 NCAA Basketball Tournament <i>Playground</i> · 10 days to go	Swag 442 teams
	Transfer Learning on Stack Exchange Tags Predict tags from models trained on unrelated topics <i>Playground</i> · 13 hours to go	375 teams
	Titanic: Machine Learning from Disaster Start here! Predict survival on the Titanic and get familiar with ML basics <i>Getting Started</i> · 3 years to go · Entered	6,109 teams
	House Prices: Advanced Regression Techniques Predict sales prices and practice feature engineering, RFs, and gradient boosting <i>Getting Started</i> · 3 years to go	2,065 teams
	Digit Recognizer Learn computer vision fundamentals with the famous MNIST data <i>Getting Started</i> · 3 years to go	1,515 teams

競賽區

徵才區

練功區

入門區

Kaggle — 數據分析競賽平台

- **2010/4創立**
- 是一個**數據建模**和**數據分析**競賽平台
- 企業和研究者可在其上發布數據，統計學者和數據挖掘專家可在其上進行競賽以產生最好的模型

Kaggle 的模組

- **Competitions** 競賽

- 獎金競賽、*Recruitment*(徵才)、*Playground*(練習)、*Getting Started*(入門)

- **Datasets** 數據集

- 提供 *Open Data* 下載、資料分析 *Code* 及論壇

- **Kernel** 核心

- 能實作並分享所有數據科學工作的平台，包括與本地工具的結合、團隊間的私有合作空間

- **Discussion** 論壇

- 任何問題都可在論壇提出，也可在此回覆問題










- **Jobs** 求職

- 資料分析相

Kaggle的排名

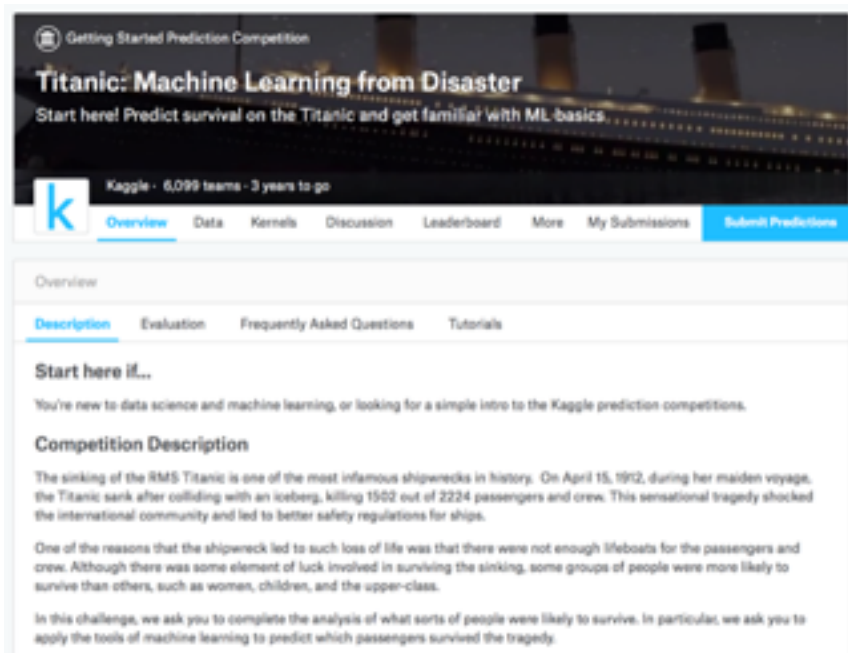
■ 排名系統

- 等級由低至高分別是*Novice*、*Contributor*、*Expert*、*Master*和*Grandmaster*
- *Competitions*中完成指定類型的比賽、在*Kernels*中分享代碼、在*Forum*中回答他人的問題都可以累計積分

Competitions Contributor	Kernels Contributor	Discussion Contributor
Unranked	Unranked	Unranked
 0	 0	 0
 0	 0	 0
 0	 0	 0
Expedia Hotel Recomm... 9 months ago · Top 67%	No kernel results	No discussion results
1,322 nd of 1974		
Santander Customer S... a year ago · Top 55%		
2,790 th of 5123		
Titanic: Machine Learn... 3 years to go · Top 49%		
2,939 th of 6098		

Getting Started – Titanic

■ <https://www.kaggle.com/c/titanic>



The screenshot shows the Kaggle 'Titanic: Machine Learning from Disaster' competition page. At the top, it says 'Getting Started Prediction Competition' and 'Titanic: Machine Learning from Disaster'. Below that, it says 'Start here! Predict survival on the Titanic and get familiar with ML basics.' and 'Kaggle · 6,099 teams · 3 years to go'. The navigation bar includes 'Overview', 'Data', 'Kernels', 'Discussion', 'Leaderboard', 'More', 'My Submissions', and 'Submit Predictions'. The 'Overview' section is active, showing tabs for 'Description', 'Evaluation', 'Frequently Asked Questions', and 'Tutorials'. The 'Description' tab is selected, displaying the text: 'Start here if... You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions. Competition Description The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.'