高中名校學生與大學表現優劣之相關性分析

摘要

本研究為「高中名校學生與大學表現優劣之相關性分析」，目的在於找出高中名校畢業學生在中央大學部分系所的學業表現，也希望想出一套方法能夠精確地表達兩者的相關性。針對我們要研究的學生在其系所內之成績進行正規化、嘗試透過數種方法加以分析、並觀察名校畢業學生與大學成績優良的相關性。

研究目的

本專案研究目的在於

1. 建立成績相關性分析之流程與實作方法
2. 透過各項方法驗證「名校學生在大學成績表現不如預期」之真實性

研究方法

- 正規化方法：依據課堂代碼分組，個別做以下兩種處理方式
  - 成績 PR 值：將排名轉換為百分位數
  - Z-score：將分數標準化 (將原分數標準差變為一，原平均變為零)
- 工具介紹：本專案採用五種分析工具
  - Survival and Probability Density Functions：通過基本統計量判斷資料特性
  - K-Means 分類器：此方法希望找出 K 組平均 μ_i 使得以下能量最小

$$E(\mu) = \sum_{i}^{k} \int_{S_i} \|z - \mu_i\|^2 \, dp(z).$$

  - Logistic Regression：為單層神經網路加上 σ 生成函數之回歸方法

$$\sigma(f(x)) = \frac{1}{1 + e^{-w(x)}} = \frac{e^{w(x)}}{1 + e^{w(x)}}.$$

  - Decision Tree：以特徵之數值分割將資料區隔，可透過 Gini impurity 判斷分類錯誤評估

$$I_G(t) = \sum_{i=1}^{C} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{C} p(i|t)^2$$

  - Random Forest：多個決策樹組合而成，可一定程度提升精準度

$$\hat{f} = \frac{1}{T} \sum_{i=1}^{T} f_i.$$

研究流程

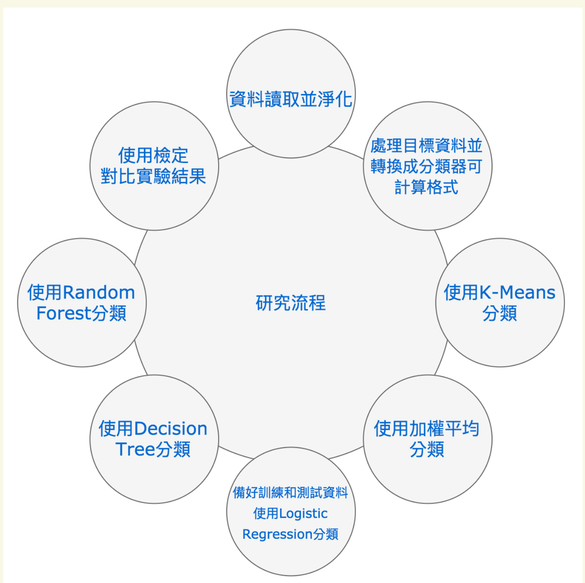


Figure: 研究流程

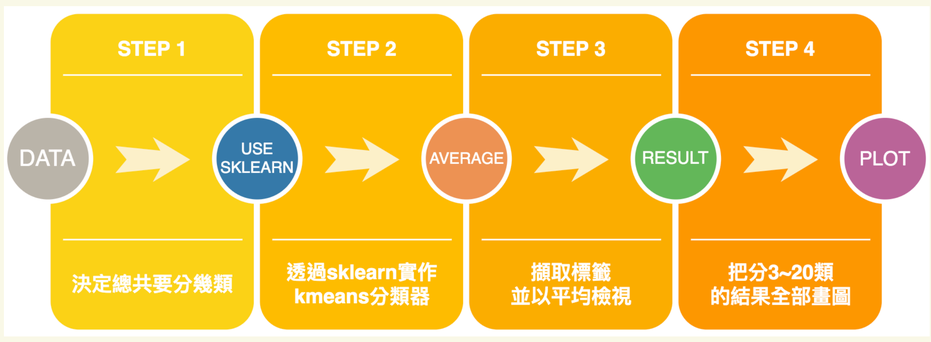


Figure: K-Mean 實作流程

研究成果

- 模組化的 Python 套件

我們建立一個模組化的 Python 套件，簡化分析校務成績之於自定義指標之相關性問題。

- 與名校成績之相關性評估

我們從兩項成績指標，分別套用五種評定方法，最終得知以下兩者結論

1． 成績與高中名校與否並不存在顯著關係
2． 從生存函數與機率密度函數中發現名校學生分布上相對非名校有極端現象

參考資料

1. J. MacQUEEN (1967). *Some Methods for Classification and Analysis of Multivariate Observations.* Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 281-297.
2. L. Breiman (2001). *Random Forests.* Machine Learning, 45(1), 5-32.
3. Stuart J. Russell and Peter Norvig (2009). *Artificial Intelligence: A Modern Approach 3rd.* Prentice Hall Press Upper Saddle River, NJ, USA.

成果展示 (樣本空間以某系為例)

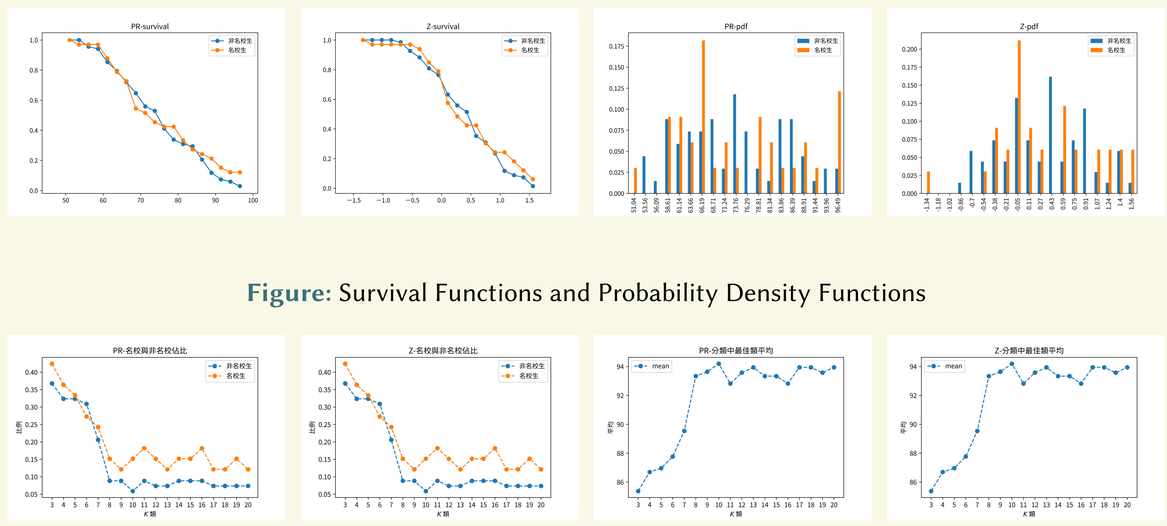


Figure: Survival Functions and Probability Density Functions

Figure: K-Means Classifier (3-20 clusters)

Logistic Regression

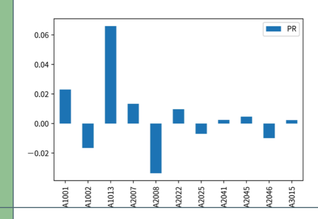
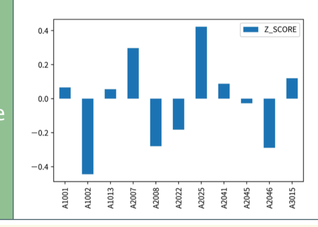
	Coefficient	Confusion matrix	Precision	Recall
PR		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 17 / 3 預測類別 名校學生: 4 / 2	0.3333	0.4
Z Score		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 19 / 1 預測類別 名校學生: 6 / 0	0.0	0.0

Figure: Logistic Regression

Decision Tree

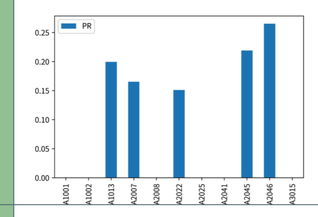
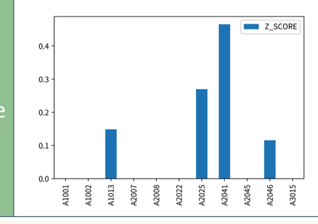
	Feature importances	Confusion matrix	Precision	Recall
PR		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 13 / 5 預測類別 名校學生: 7 / 1	0.125	0.166
Z Score		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 17 / 1 預測類別 名校學生: 7 / 1	0.125	0.5

Figure: Decision Tree

Random Foerest

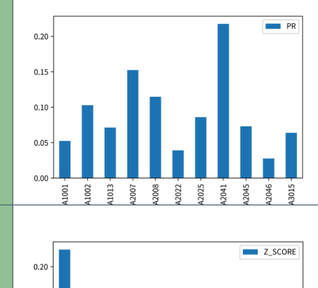
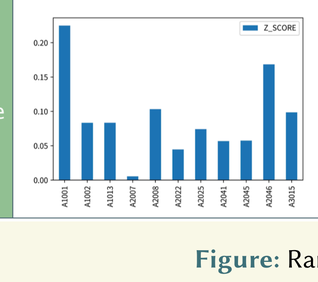
	Feature importances	Confusion matrix	Precision	Recall
PR		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 14 / 1 預測類別 名校學生: 9 / 2	0.18181	0.666
Z Score		實際類別 / 非名校學生 / 名校學生 預測類別 非名校學生: 15 / 0 預測類別 名校學生: 10 / 1	0.0909	1.0

Figure: Random Forest

未來工作

- 更合理的資料淨化方法：建立轉系、抵修、被二退學生之成績缺失值處理方法
- 擴展樣本空間：針對更多系所進行分析
- 擴展分析人數：放寬學生選取條件

生薑筍
彭琍棱、黃冠澤、林育愷、李家妤