# 數據科學分析的土壤一 Trust AI & Data Quality Pipeline

專題學生:周成康、陳瑾叡、吳驊祐、史康宇、林孟璇、黃戎僔 指導教授:李家岩

指導助教:陳彥彣、方鈺學

#### Study Background

近期,大型語言模型引領了新的模型熱潮,在強化 模型表現之餘,模型的可靠性、資料品質也同等重要, 它們能大幅影響訓練成本、泛化性及易用性。因此我們圍繞 Trust Al 以及 Data Quality 雨大主題,建立了一個數據 流程平台,為使用者提供數據處理與分析的工具。透過多 程序的資料推論,建立自動化的分析流程,以節省資料 成本,並提升AI模型預測可靠度。

Trust AI 能解釋預測結果並提供可靠性,提升泛化能力。 SHAP 能評估資料的因子如何影響預測,知識蒸餾能將複雜 模型壓縮, Active learning 則提供動態標記並訓練重點資料。

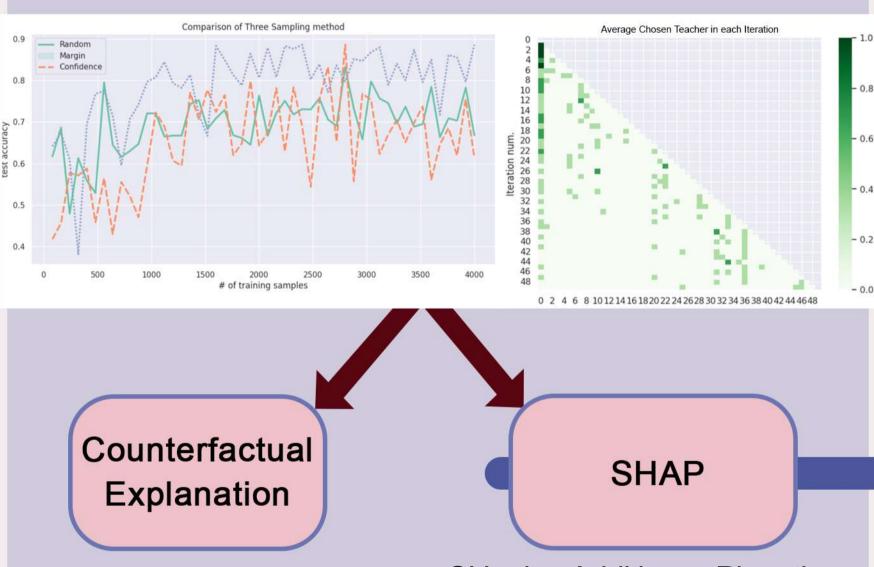
Data Quality能大幅影響模型結果的好壞,資料缺失、無用資訊會造成「垃圾進,垃圾出」的情形。我們著眼於各情 境資料補值的方法論,及資訊量評估指標,透過客觀的量測

### 指標,評估資料處理的變化,維持資料完整性及資訊量。 Flow Chart 原始、有缺失 的資料 五大方法: ·Fuzzy C-means 進行補值 ·Missing Forest ·KNN VAE 進行指標計算 Multi-Differential JS-**NRMSE Entropy** Collinearity Divergence ▼ The distribution and similarity of data imputed before/after 系統根據資料取得策略,在每輪 補值完的資料進行 次中取得對當下最有幫助的未標 **Active Learning** 注資料加進訓練集train,直到使 用者滿意模型表現。

Input:Teacher model Knowledge Distillation Output:Student model

知識蒸餾可把模型輸出之預測機率平滑 化,得軟標籤後結合資料Ground Truth (1) 自部分過去迭代時的訓練模型學習 (2) 將大模型的知識以及泛化能力遷移 到較好攜帶、部署的小模型。

Reference: Trustworthy Active Learning Using Knowledge Distillation (Beong-woo Kwak, 2022) 上述過程使用驗證集尋找能與當下模型相輔相成的過往模型 協同訓練,目標是得到乾淨有價值之訓練集,減少人工標記耗損。



若SHAP不足以提供適當決策 輔助方案,可使用反事實解釋 (Counterfactual Explanation) 。其實例呈現機器學習模型輸 入值之變化如何影響預測結果 ,直觀、直接地告知使用者如 何調整輸入以達到理想結果。

SHapley Additive exPlanations 是Trust AI 領域的一項技術,可 計算特徵對模型預測的邊際貢獻 ,解釋模型預測值為每個特徵歸 因值的總和,協助理解模型預測 的原理。模型訓練後導入此技術 可增強模型的解釋性和可信度。

## EDASH - Data Imputer & Data Qaulity Metric

#### **Data Imputation**

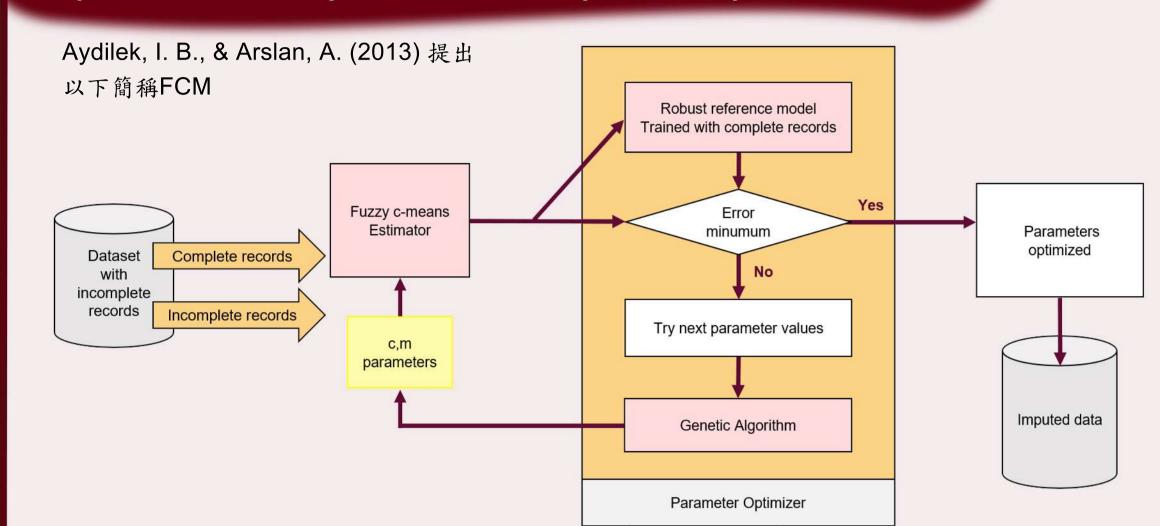
特定領域如測驗型問卷或稀缺資料,常容易有資料缺失的情況,此步驟針對 前處理的資料缺失值進行填補,為多種不同情境提供適合的 Imputer,如連續、 類別型資料,或有、無母數等假設。我們研究相關文獻並實作出比一般補值方式 表現更好的方法,包括 ML/DL-based,如 Optimized Fuzzy C-means, Expectation-Maximization 以及 MissForest 等。

#### **Data Quality Metric**

為評估資料品質的性質,並配合 Imputer 觀察資訊量變化,除了敘述性統計 指標與類別型資料的 Entropy-Based Metric,也針對連續型資料設計 Differential-Entropy Based Metric 以補足對於資訊量變化的評估方法,如 Differential JS-Divergence。使類別及連續型資料都有完整的資料品質評估標準。

## EDASH - Data Imputer 資料補值:

#### Optimized Fuzzy C-means Imputer Implementation

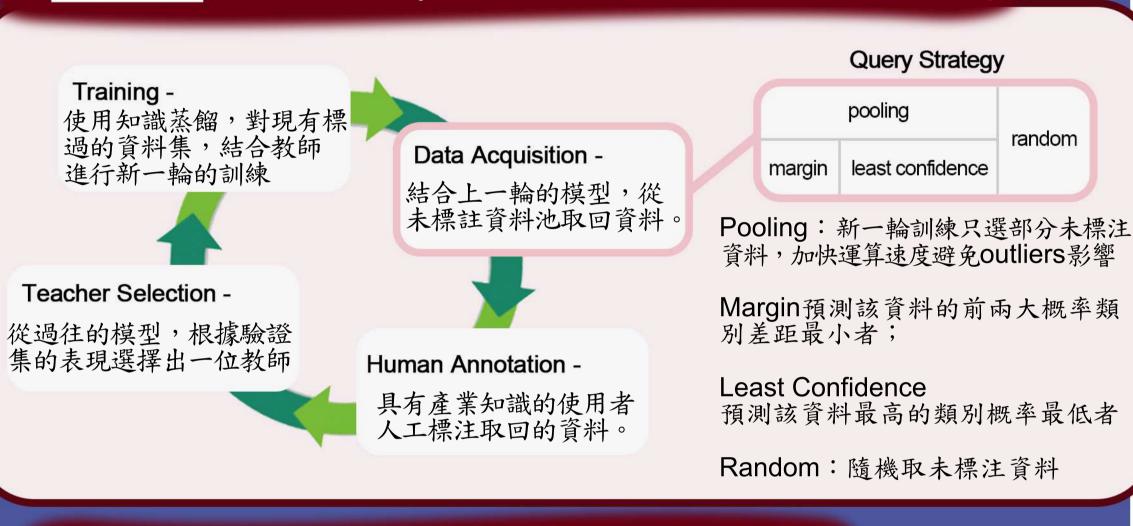


此 Multi-model 的補值流程被用於更新 Main Model,即 FCM 的 Weighting Factor 和 Cluster Number。FCM 具有無母數及Soft-clustering的特性,對於不 穩定的訓練情境包容度較佳。

將經過 Complete Record 訓練的 Robust Reference Model Ouput 與 FCM的Output 進行比較,以達到正則化,並防止 Overfitting,強化模型表現。

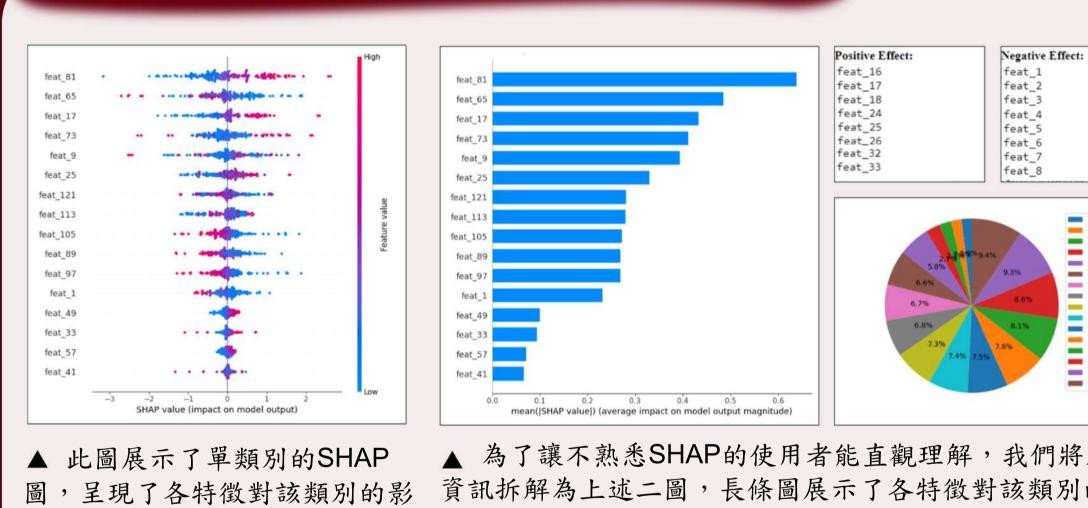
參數會再經過基因演算法最佳化後,交給 FCM 再預測,流程依據使用者需 求持續或終止,如設定最小接受域或遞迴次數。

## Trust AI Module Implementation 1: Active Learning



# **Trust AI** Module Implementation 2: SHAP

表示對目標值的正向影響。

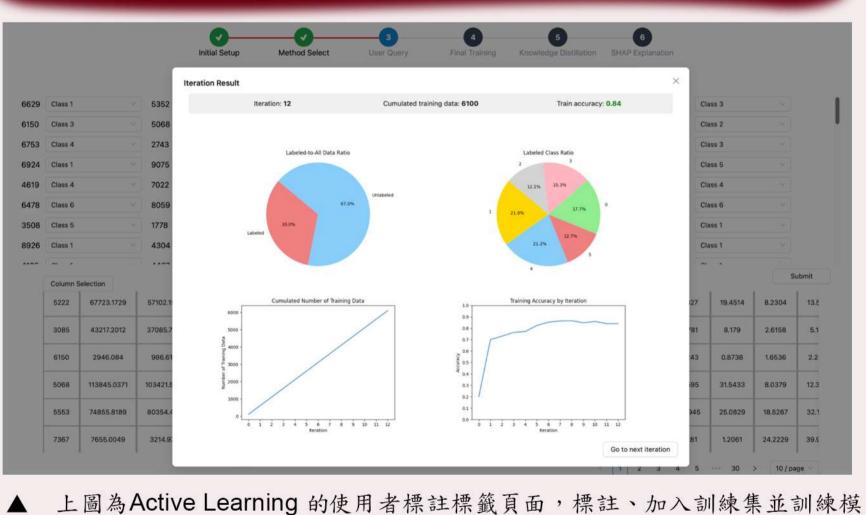


▲ 為了讓不熟悉SHAP的使用者能直觀理解,我們將左圖 資訊拆解為上述二圖,長條圖展示了各特徵對該類別的影 響大小。文本欄則顯示對目標值有正負向影響的特徵。圓 響,特徵的資料點分布寬度反映 餅圖除了每個特徵的影響外,也顯示選擇各特徵能夠涵蓋 其影響程度,而中線右側的紅色 之百分比資訊量,協助決策該類別的預測。 feat 1 feat 2 feat 3 feat 4 feat 5 feat 6 feat 7 feat 8 feat 9 feat 10 feat 11 feat 12 feat 13 feat 14 feat 15 feat 16 feat 17 Gas Class

20.2 25.734 36.077 -104.8 126042.2 7.9056 26.782 34.512 47.654 20.2 25.734 36.077 -104.8 126042.2 7.9056 26.782 34.512 47.654

最後Counterfactual Explanation 協助決策。表中第一行為輸入資料,後兩行展示建議修改方案。

# Case Study 案例探討- Gas Sensor Dataset



型後,會顯示出該迭代結果與累積數據,如過去所有迭代的訓練準確度之線圖等。

此實驗採用Gas Sensor資料集,並統一採用同個模型做分類。完整資料的F1 Score為 0.9928,下圖為各方法與不同缺失率的F1 score比較表。可發現資料本身特徵良好,即便在缺失 25%資料的情況,使用均值填補,都能取得一定表現,但在缺失率提升的情況下,我們研究的 方法逐漸拉開均值、KNN的表現。即便在75%缺失率,都能有接近完整資料的預測效果。而 FCM的算法為需要一定數量的完整資料點,在缺失率高的極端情況,有較明顯的表現遞減。

缺失率\補值方法	mean	knn	fuzzy	em	missforest
25%	0.9861	0.9875	0.9866	0.9908	0.9914
50%	0.9573	0.9383	0.9547	0.9899	0.9894
75%	0.8573	0.8714	0.8623	0.9742	0.9800

結論而言,Imputers 表現優於常見方法,缺失25%時,對比完整資料僅不到1%準確性下降。 其次,Active Learning 提供以使用者意見更新資料 Label 的介面,有效降低訓練成本。知識蒸餾 壓縮參數量之20%,準確性僅下降 2%,能節省運算及部署成本。最終,SHAP解釋資料點因子 不同分佈於預測結果(Y)之影響,配合反事實解釋,進行反向推論,協助決策輔助。此平臺強調 模型訓練中易被忽視但同等重要的部分,期望能使資料科學的應用更具解釋性與效率。

