National Taiwan University **Introduction to Machine Learning and Deep Learning**
Department of Civil Engineering Instructor: C.-S. CHEN

Midterm Report
Due 23:59, Tuesday, October 31, 2023

Student ID: R11522659
Name: 卓靖鎧

---

1. 請說明你如何進行資料前處理。**(1%)**

1. **刪除同學請教土木系朋友哪些 feature 是在評估校舍耐震能力不會使用到的，以及一些數據過少的欄位（我設定小於 500 筆數據的欄位需刪除），總共 8 欄**
2. **將年代過大、過小的數值進行 UNIX 時間調整，以及將所有欄為小於 0 的數改為 0，超過數據分布 90%的 10 倍改為 NAN**
3. **透過 IterativeImputer 及 BayesianRidge 模型將缺失的數據進行補足，再透過 RobustScaler 對資料的分布進行正規化**
4. **先用 XGBoost 找出 important feature 中排名前 8 個，再用這 8 個欄位的資料進行 Smote，讓 label 的分布平均**

2. 請說明你所使用的 **model** 以及 **hyper-parameter tuning** 的心得。**(1%)**

我最後是使用 XGBoost，我有試過其他的模型，像是 LightBGM、Random Forest、CatBoost 等等，但效果似乎都沒有 XGBoost 好，不過我認為最主要的應該是我沒有將其他模型的參數調好，才導致其他模型的結果不好。

在調整 hyper-parameter 的部分，我其實都只有在調整模型的深度、學習率以及 gamma 值，因為其他的參數我都不太了解，調整之後 train 出來的結果變化也不大，而且參數過多的話，我根本調不完。雖然我有使用 GridSearchCV，但我覺得它的效果有限，就算在 train 的數據找出最適合的參數，但預測 X_test 的結果通常沒有比較好，而且只要放入 GridSearchCV 的參數稍微多一點，計算的時間是倍數成長，所以我最後都直接將結果放到 Kaggle 看分數有沒有比較好。

最後我發覺調參數似乎沒有任何依據，我有分別試了兩種不同的深度，一個深度是 9，另一個是 50，結果上傳到 Kaggle 的分數居然都差不多在 0.803 附近，試完這兩個參數後，我就認為模型參數雖然重要，但更重要的一定是資料前處理，不然深度差這麼多，怎麼結果差不多。

3. 畫出 **confusion matrix** 分析 **model** 分類的結果，並列出 **precision**、**recall** 和 **F1-score**，再加以簡單說明。**(2%)**

我將經過第一點（前處理）的資料切分成 train : 80%、validation : 20%，train 的拿來 fit XGBoost 模型，validation 拿來驗證模型。因為經過 Smote，所以資料總數為 5328 筆，train 的資料個數：4262、test 的資料個數：1066

	PREDICTED CLASS		
ACTUAL CLASS		Yes (0)	No (1)
	Yes (0)	TP = 456	FN = 58
	No (1)	FP = 84	TN = 468

Results：

1. $Yes : No = 456 + 58 : 84 + 468 = 514 : 552 \approx 1 : 1$
2. $Accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{456+468}{456+84+58+468} = 0.8667$
3. $Precision = \frac{TP}{TP+FP} = \frac{456}{456+84} = 0.8444$
4. $Recall = \frac{TP}{TP+FN} = \frac{456}{456+58} = 0.887$
5. $F-msasure = \frac{2TP}{2TP+FN+FP} = \frac{456*2}{456*2+58+84} = 0.87$

結論：

在預測校舍是否穩定的情況下，我認為最重要的是是避免將不耐震的校舍判斷為耐震的，也就是 FP 要盡可能的低，因為這對學生來說非常危險，而我的模型只有 84%的精準度能夠判斷正確，這實際上是非常需要加強的。

相較之下，在實際穩定的校舍中，如果預測出不穩定的話也就只是在多加一層防護，其實對學生來說影響較小，而我的模型有能夠判斷出穩定校舍中的 89%，雖然還可以再提高準確率，但我認為這不是模型最需要先改善的部分。

Submission Format

Convert `midterm_report_template.docx` to `midterm_report.pdf`, then place `midterm_report.pdf` and `codes` into a folder named {yourStudentID}_midterm and compress it into a ZIP file for upload to NTU COOL. Below is the file format example for upload.