USING REGRESSION ANALYSIS TO PREDICT ADULT WORKING HOURS PER WEEK AND AIRFOIL SELF-NOISE DECIBELS

以迴歸分析模型來預測成人一週工作時數以及機翼自噪分貝

林郁凱，彭冠穎，彭嘉瑋，陳羿欣
Department of Information Management, National Yunlin University of Science and Technology

摘要

本研究利用隨機樹(Random forest)、XGBoost、支援向量迴歸(Support Vector Regression，SVR)以及類神經網路(Neural network)不同的迴歸分析模型來預測機翼自噪分貝與成人一週的工作時數，並以 MAPE 及 RMSE 作模型的績效評估，最終探討兩組資料集在不同演算法的預測績效。

關鍵字：隨機樹、XGBoost、支援向量迴歸、類神經網路、迴歸分析、MAPE、RMSE。

一、緒論

1.1 動機

本研究利用四種不同的數值預測模型來進行兩組資料集的迴歸分析；數值預測模型分別為隨機樹(Random forest)、XGBoost、支援向量迴歸(Support Vector Regression，SVR)以及類神經網路(Neural network)，以上四個模型共同之處為能夠用來處理連續的預測問題，本研究主要探討機翼自噪分貝數據集（Airfoil Self-Noise Data Set）與成人資料集（Adult Data Set）利用上述四種不同的預測模型進行數值預測，以 MAPE 及 RMSE 作為模型的評估指標，並探討兩組資料集在不同演算法的預測績效[1][2]。

迴歸分析(Regression Analysis)是機器學習中最常見的應用之一，其方法主要探討變數與變數之間的相關性，透過迴歸模型可以來推論預期的結果，在本研究中使用自噪分貝數據集（Airfoil Self-Noise Data Set）來預測機翼葉片自流時產生的噪聲分貝以及成人資料集（Adult Data Set）預測成人一週的工作時數。

1.2 目的

迴歸問題的準確性，需要建立一個評估迴歸模型擬合效果的指標，因此在本實驗的最後會以平均絕對百分比誤差(Mean absolute percentage error，MAPE)以及均方根誤差(Root Mean Squared Error，RMSE)比較模型績效。

二、方法

2.1 實作說明

在迴歸分析模型的評估實驗中，本研究首先將「機翼自噪分貝數據集（Airfoil Self-Noise Data Set）」、「成人資料集（Adult Data Set）」做數據的前置處理，其包括資料清理、Ono-hot encoding、資料切割（即為將資料分成訓練資料(train data)以及測試資料(test data)），並使用隨機樹(Random forest)、XGBoost、支援向量迴歸(Support Vector Regression，SVR)、類神經網路(Neural network)，四種不同的數值預測模型，以 MAPE 及 RMSE 做模型的績效評估。

2.2 操作說明

本研究執行環境皆為 Python3.6，以 Anaconda Jupyter Notebook 作為分析工具，利用 Pandas、Numpy 來讀取資料以及做資料的前處理，預測模型則利用 Scikit - learn 套件來建構，最後再將測試資料導入，並且顯示其預測數值。

三、實驗

3.1 資料集

本研究使用兩組資料集做預測分析，分別為機翼自噪分貝與成人收入調查結果之相關數據，以下為該兩組資料集之資料名稱、資料筆數，以及資料表的欄位介紹。

3.1.1 機翼自噪分貝數據集（Airfoil Self-Noise Data Set）

機翼自噪分貝數據集為美國 NASA 提供之 NACA 0012 機翼的各種風動的速度、迎角等數據，該資料集之數據收集中，觀察者的位置與機翼的跨度為固定的，而其跨度(span)即為機翼尖端至另一機翼尖端的距離[2][3]。

- 名稱：Airfoil self-noise 資料集
- 原始資料筆數：1503
- 正規化後之訓練資料筆數：1,202
- 正規化後之測試資料筆數：301

表一：Airfoil self-noise 資料集欄位介紹

欄位	欄位名稱	資料屬性	單位
0	Frequency	Continuous	赫茲(Hertz)
1	Angle of attack	Continuous	角度(Degrees)
2	Chord length	Continuous	公尺(Meters)
3	Free-stream velocity	Continuous	公尺/秒(Meters per second)
4	Suction side displacement thickness	Continuous	公尺(Meters)
5	Scaled sound pressure level	Continuous	分貝(Decibels)

表二： 顯示部分 Airfoil self-noise 資料集(前五筆與最後五筆)

（欄位編號對應表一之欄位名稱）

名稱 / 編號	Frequency	Angle of attack	Chord length	Free-stream velocity	Suction side displacement thickness	Scaled sound pressure level
0	800	0	0.3048	71.3	0.00266337	126.201
1	1000	0	0.3048	71.3	0.00266337	125.201
2	1250	0	0.3048	71.3	0.00266337	125.951
3	1600	0	0.3048	71.3	0.00266337	127.591
4	2000	0	0.3048	71.3	0.00266337	127.461
…	…	…	…	…	…	…
1498	2500	15.6	0.1016	39.6	0.052849	110.264
1499	3150	0	0.3048	71.3	0.00266337	125.201
1500	4000	15.6	0.1016	39.6	0.052849	106.604
1501	5000	15.6	0.1016	39.6	0.052849	106.224
1502	6300	15.6	0.1016	39.6	0.052849	104.204

3.1.2 成人資料集（Adult Data Set）

成人資料集為 Barry Becker 從 1994 年人口普查數據庫中收集而成[1]，該資料集的主要任務為分類成人年薪資是否為 5 萬元，但在本研究之實驗，將之應用於迴歸分析之任務上，並且預測成人一週的工作時數。

- 名稱：Adult 資料集
- 原始資料筆數：48842
- 正規化後之訓練資料筆數：30162
- 正規化後之測試資料筆數：15059

表三 ：Adult 資料集欄位介紹

欄位	屬性	內容
0	Age	continuous
1	workplace	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
2	Fnlwt	continuous
3	education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
4	education-num	continuous
5	marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
6	occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
7	relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
8	Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
9	Sex	Female, Male

續下頁

承上頁		
欄位	屬性	內容
10	capital-gain	continuous
11	capital-loss	continuous
12	hours-per-week	continuous
13	native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
14	salary	<=50K,>50K

表格四：顯示部分 Adult 資料集

編號 / 欄位名稱	0	1	2	3	4
age	39	50	38	53	28
workplace	State-gov	Self-emp-not-inc	Private	Private	Private
fnlwt	77516	83311	215646	234721	338409
education	Bachelors	Bachelors	HS-grad	11th	Bachelors
education-num	13	13	9	7	13
marital-status	Never-married	Married-civ-spouse	Divorced	Married-civ-spouse	Married-civ-spouse
occupation	Adm-clerical	Exec-managerial	Handlers-cleaners	Handlers-cleaners	Prof-specialty
relationship	Not-in-family	Husband	Not-in-family	Husband	Wife
race	White	White	White	Black	Black
sex	Male	Male	Male	Male	Female
capital-gain	2174	0	0	0	0
capital-loss	0	0	0	0	0
hours-per-week	40	13	40	40	40
native-country	United-States	United-States	United-States	United-States	Cuba
salary	<=50K	<=50K	<=50K	<=50K	<=50K

3.2 前置處理

圖 1 前置處理流程圖

- 資料清理：將資料中的缺失值以及過濾與預測結果不相關資訊。
- One-Hot encoding：對非數值的類別屬性進行特徵數字化。
- 資料切割：將 80%當成訓練資料，其餘 20%為測試資料。

表格五：機翼自噪分貝數據集選擇資料集之資料前處理後部分資料

欄位名稱 / 編號	Frequency	Angle of attack	Chord length	Free-stream velocity	Suction side displacement thickness	Scaled sound pressure level
0	800	0.0	0.3048	71.3	0.002663	126.201
1	1000	0.0	0.3048	71.3	0.002663	125.201
2	1250	0.0	0.3048	71.3	0.002663	125.951

表格六：成人資料集之資料前處理後部分資料

age	education-num	hours-per-week	new_income	workclass_Federal-gov	workclass_Local-gov	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...
39	13	40	0	0	0	0	0	0	...
50	13	13	0	0	0	0	0	1	...
38	9	40	0	0	0	1	0	0	...
53	7	40	0	0	0	1	0	0	...
28	13	40	0	0	0	1	0	0	...

3.3 實驗設計

以下為機翼自噪分貝資料集與 Adult 資料集迴歸分析實驗之設計，依序的流程步驟如圖 2：

圖 2：實驗流程圖

3.3.1 機翼自噪分貝數據集

1. 設定模型初始參數
2. 將欲處理之資料分別匯入四個模型中並進行訓練
3. 將驗證資料匯入訓練完成之模型中預測並產出結果
4. 利用預測結果算出 MAPE 及 RMSE 回歸指標
5. 測試多次參數並採用最佳結果 (結果呈現於 3.4 節實驗結果)
6. 匯出預測測試資料的結果

3.3.2 部分國家成人資料

1. 設定模型初始參數
2. 將欲處理之資料分別匯入四個模型中並進行訓練
3. 將驗證資料匯入訓練完成之模型中預測並產出結果
4. 利用預測結果算出 MAPE 及 RMSE 回歸指標
5. 測試多次參數並採用最佳結果 (結果呈現於 3.4 節實驗結果)
6. 匯出預測測試資料的結果

3.4 實驗結果

以下為機翼自噪分貝資料集以及成人資料集使用四種不同迴歸分析模型之訓練結果以及績效評估：

3.4.1 SVR

由圖 3、4 可以得知在 SVR 模型中不管是 RMSE 或是 MAPE 的評估下，其 Kernel 參數 RBF(radial basis function)會比 Sigmoid 的績效好，所以本實驗選用 RBF 作為參數。

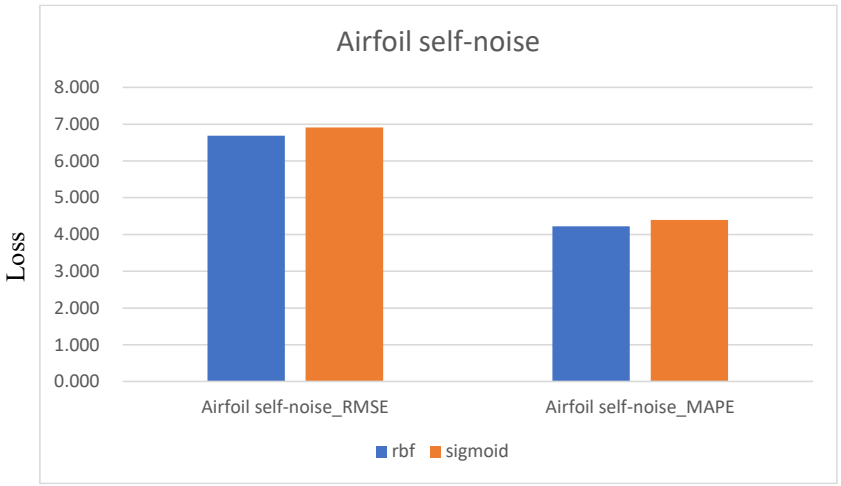

圖 3 Airfoil self-noise SVR 績效指標差異圖

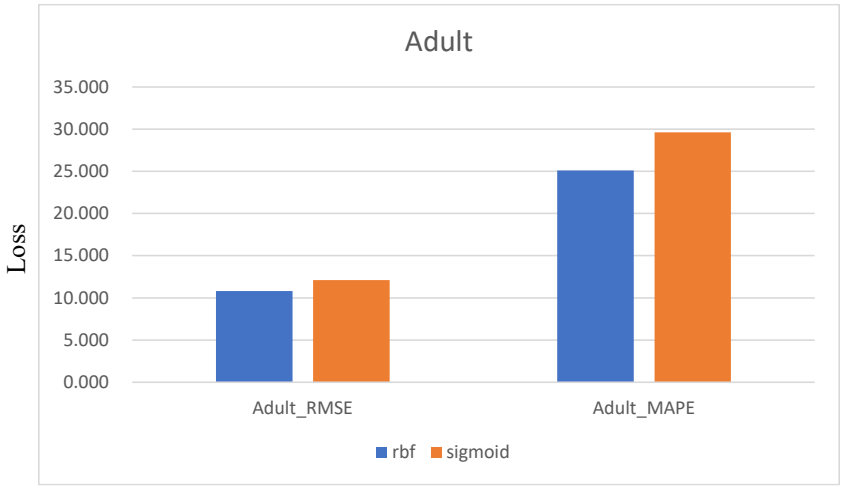

圖 4 Adult SVR 績效指標差異圖

3.4.2 類神經網路

在類神經網路模型中，以 RMSE 和 MAPE 數值最低之結果作為本實驗參數，如圖 5 及圖 6 之藍線，在機翼自噪分貝資料集中，深度參數 4,945 為 RMSE 和 MAPE 的最低點，而在成人資料集中深度參數 4,529 為 RMSE 和 MAPE 的最低點。

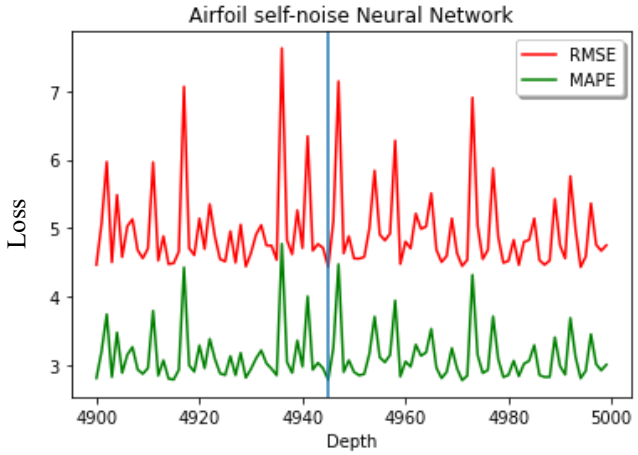

圖 5：Airfoil self-noise 類神經網路績效指標差異圖

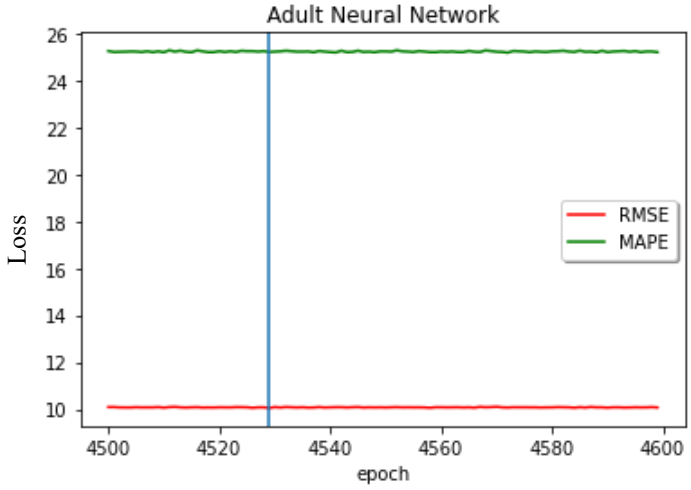

圖 6：Adult 類神經網路績效指標差異圖

3.4.3 Random Forest

圖 7 上說明在機翼自噪分貝資料集中深度參數 17 為 RMSE 和 MAPE 的最低點，所以我們選擇深度參數 17 來做隨機森林模型之訓練參數。

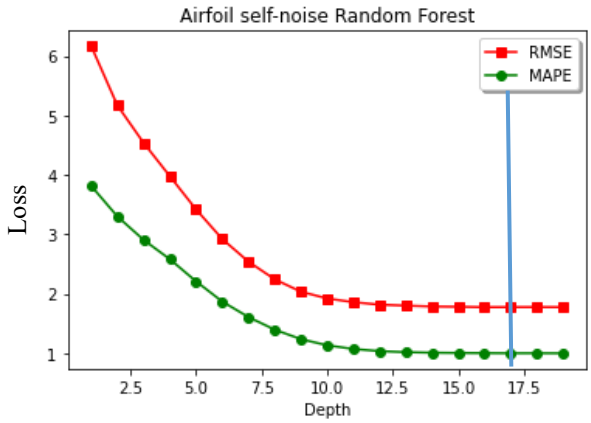

圖 7　Random forest 績效指標差異圖

在圖 8 上說明在成人資料集中深度參數 10 為 RMSE 以及 MAPE 的最低點，所以我們選擇深度參數 10 來做隨機森林模型之訓練參數。

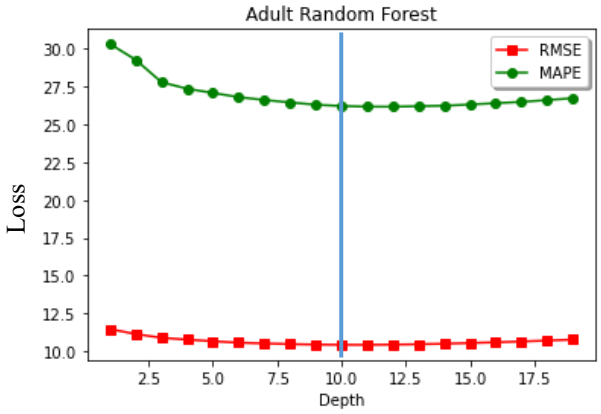

圖 8 Adult Random forest 績效指標差異圖

3.4.4 XGBoost

在 Airfoil self-noise 數據集中 Depth 在 13 時是最低點，所以我們使用 13 來做我們 depth 的參數，如圖 9 所示。

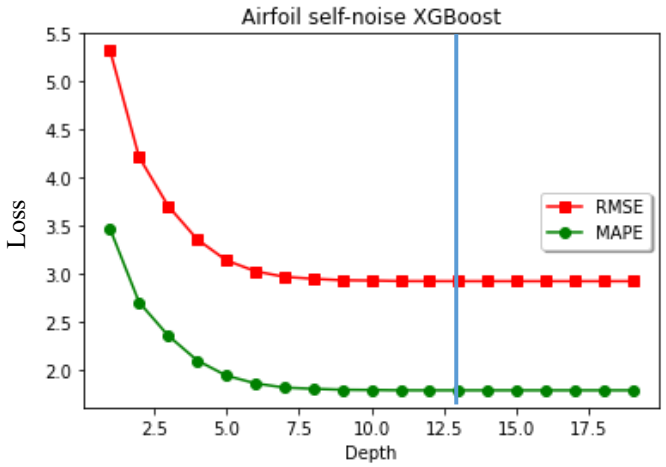

圖 9：Airfoil self-noise XGBoost 績效指標差異圖

在成人資料集中深度參數 10 時是 RMSE 和 MAPE 最低點，所以我們選擇深度參數 10 來做 XGBoost 模型之訓練參數，如圖 10 所示。

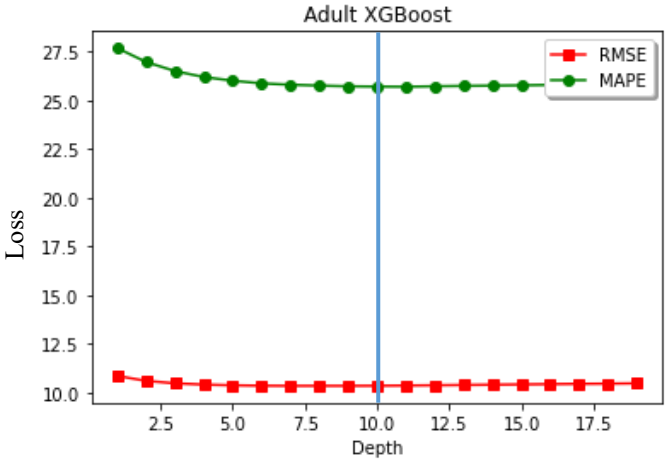

圖 10：Adult XGBoost 績效指標差異圖

四、結論

根據圖 11、12 發現在成人資料集的績效指標不論是 MAPE 或是 RMSE 都比機翼自噪分貝數據集還不佳，因此我們得知在成人資料集 hours-per-week 之欄位(一週的工作時數)的預測績效不是那麼顯著，而在機翼自噪分貝中預測分貝數值，由實驗結果得知績效指標明顯優異很多，由於成人資料集本身的特性為適合做分類之任務，因此在迴歸分析的績效較為不佳，而機翼自噪分貝資料集，由於主要被使用來做迴歸分析之用途，因此在迴歸分析中有較好的績效。

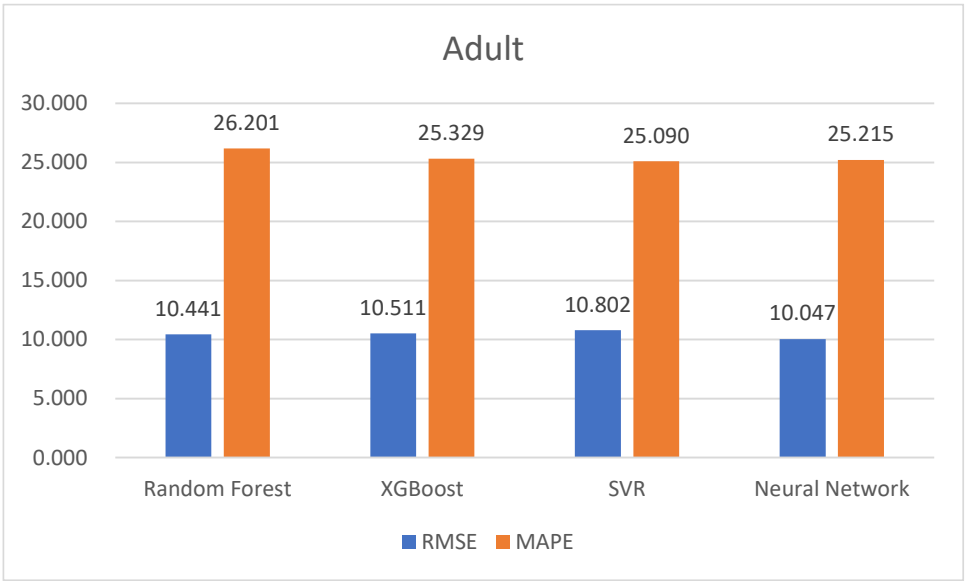

圖 11：Adult 總績效圖

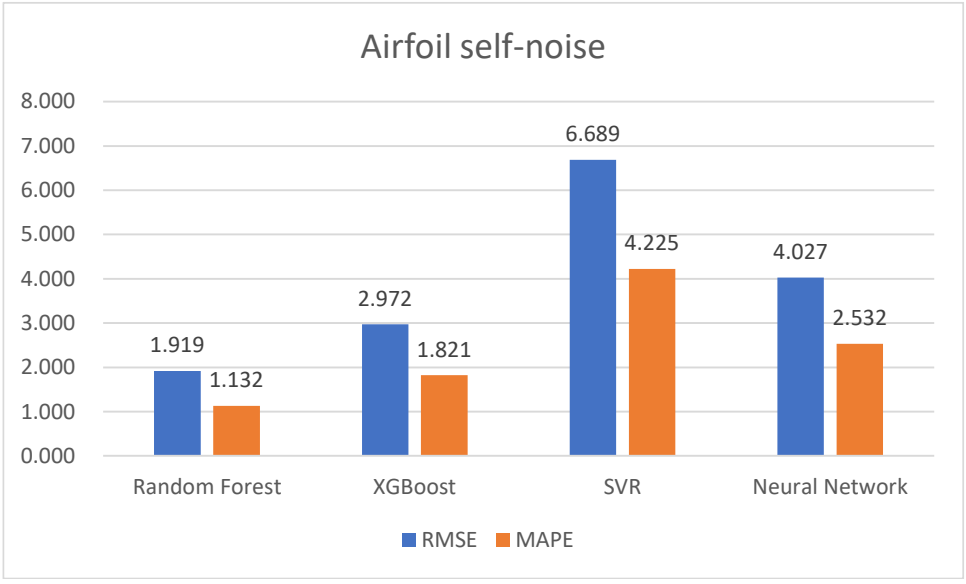

圖 12：Airfoil self-noise 總績效圖

參考文獻

[1] Tjen-Sien Lim (1997) A subset of the 1987 National Indonesia Contraceptive Prevalence Survey[Data Set].

[2] Ronny Kohavi, Barry Becker(1996) Adult Data Set[Data Set].

[3] Wing Geometry Definitions - Re-Living the Wright Way – NASA (https://wright.nasa.gov/airplane/geom.html)

[4] Post pruning decision trees with cost complexity pruning (https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html#sphx-glr-auto-examples-tree-plot-cost-complexity-pruning-py).