應用機器學習平台於衛星產品的開發

章鶴群、陳冠儒、張育承
林賢宏、羅時宜
藍秀仁、蔡宜真
楊惠婷

中央氣象局第四組/氣象衛星中心
樺鼎商業資訊股份有限公司
DataRobot台灣分公司
DataRobot新加坡分公司

DataRobot

大綱

- 應用機器學習於衛星產品的緣由
- DataRobot如何幫助衛星產品的開發
- 機器學習的結果分析討論
- 心得與總結及後續

緣由

氣象站的觀測：人工觀測→自動化觀測

可以由電子儀器取代：溫度、壓力、溼度、雨量...

難以由電子儀器取代：能見度、天氣現象、雲量、雲狀...

嘗試解決的方法：**利用同步衛星資料反演** But...

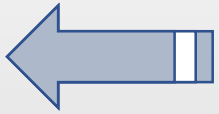
緣由

以衛星資料反演
優點：

- 無氣象站的地區亦可以得到觀測資料。
- 高時間解析度（10分鐘）優於人工觀測。

缺點：

- 易受雲的干擾
  - 例如：有雲的地方難判斷雲底之下是否有霧
- 衛星視角與觀測員的視角差異大
  - 例如：雲量的判斷，觀測員和衛星看的方向不同

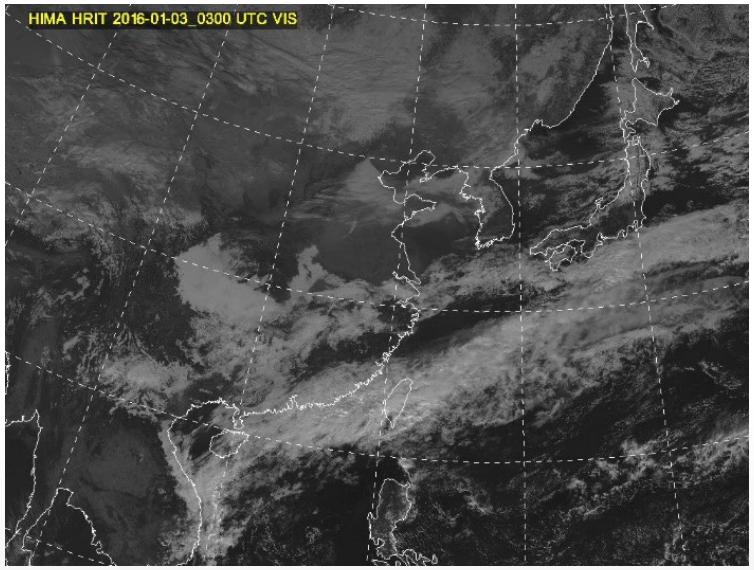


機器學習是否能夠解決這些問題？

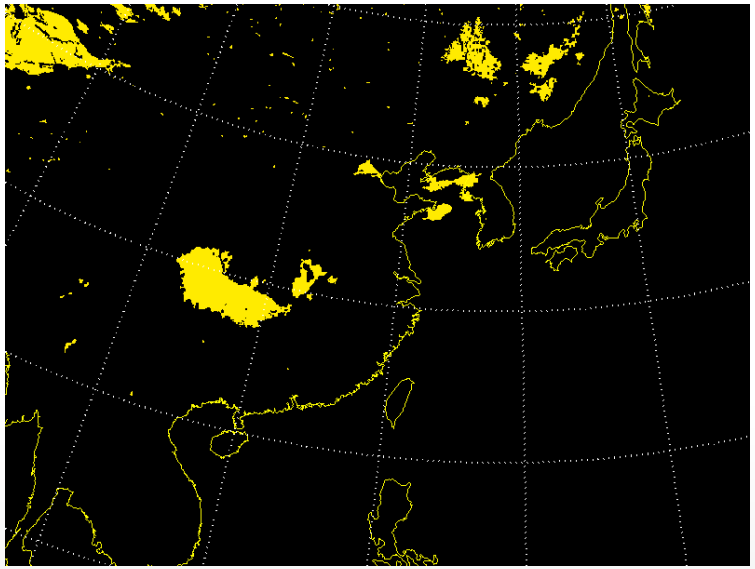
霧

頻道	波段
3	0.64
7	3.9 （短波紅外線）
14	11.2 （長波紅外線）

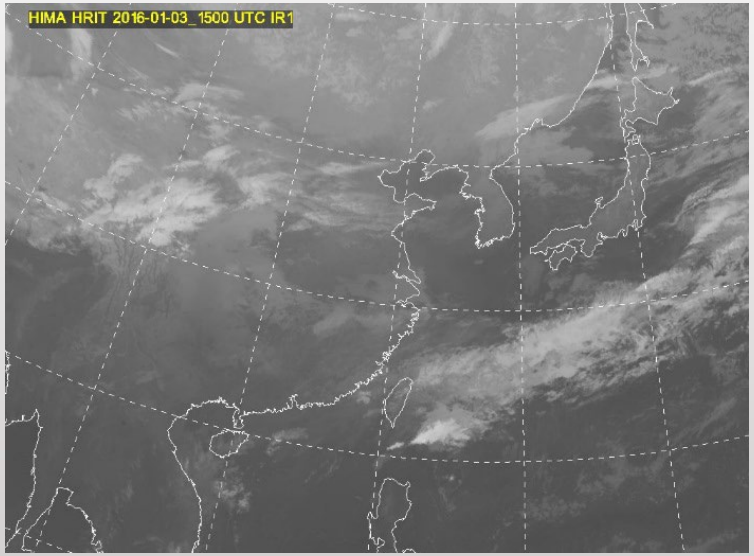
可見光雲圖



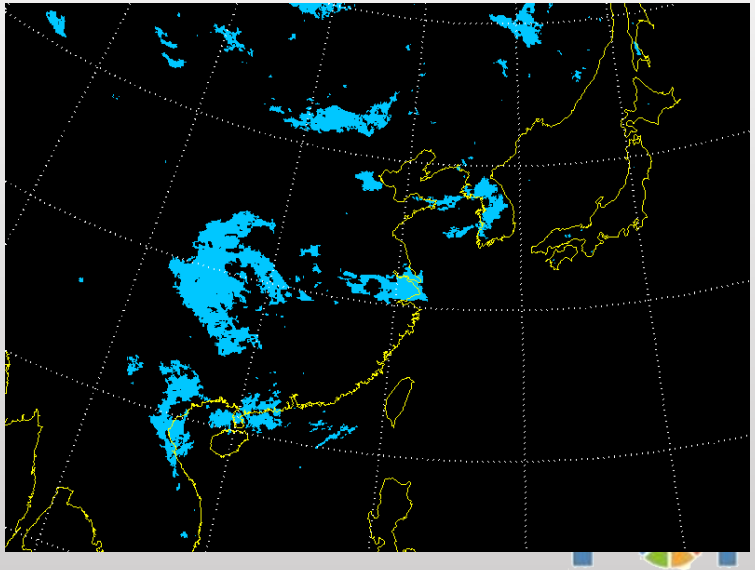
低雲與霧（日間）



紅外線雲圖

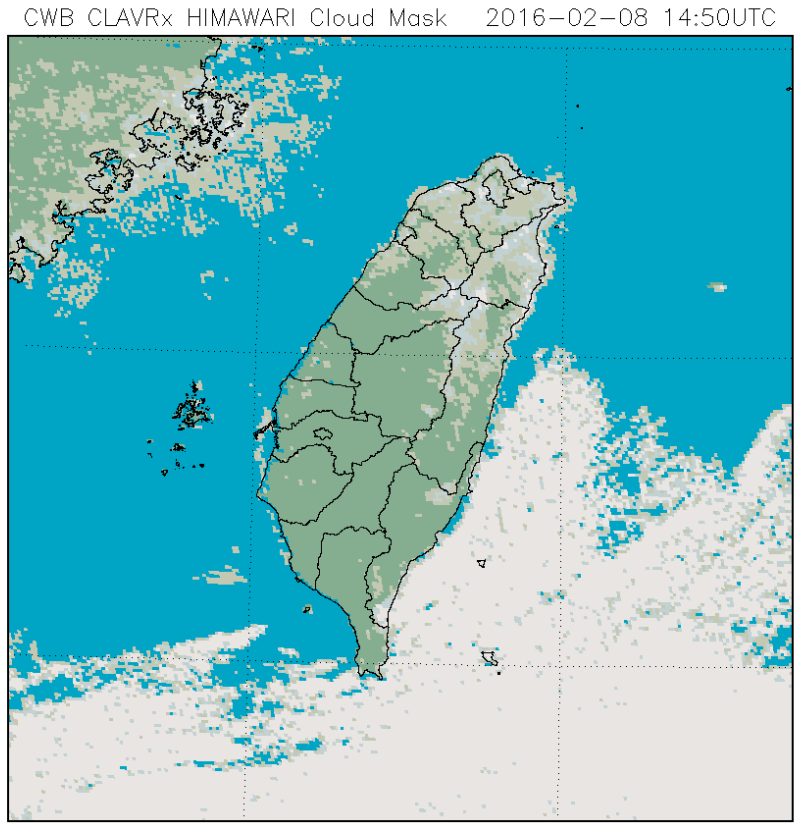


低雲與霧（夜間）

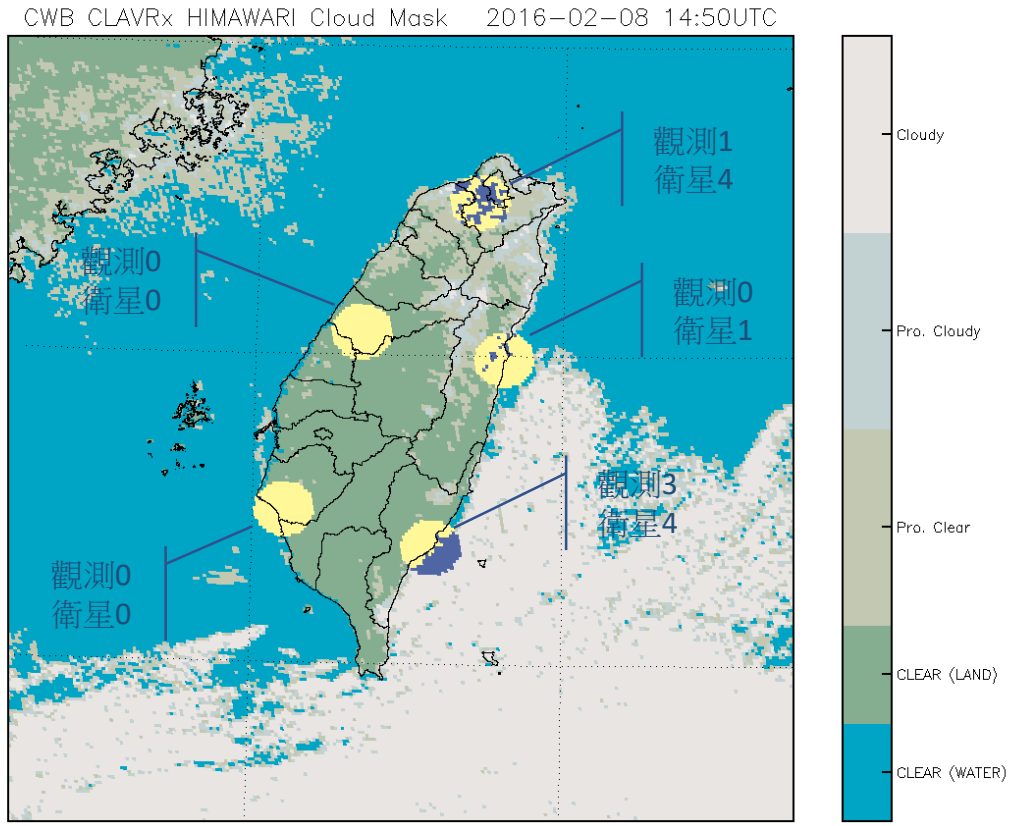


雲量

雲遮Cloud Mask



雲量



DataRobot 大量降低技術門檻

使用者無需具備

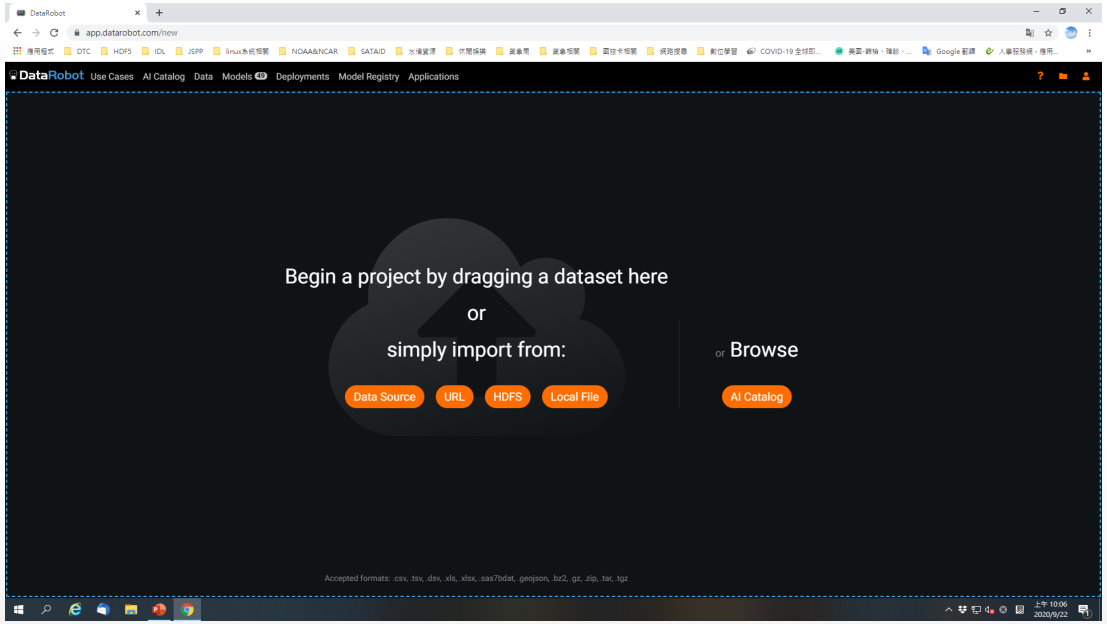
- 程式開發能力
- 進階的ETL資料處理技術
- 高深數理理論

DataRobot 進行自動化機器學習

透過十個步驟自動化整個建模生命週期

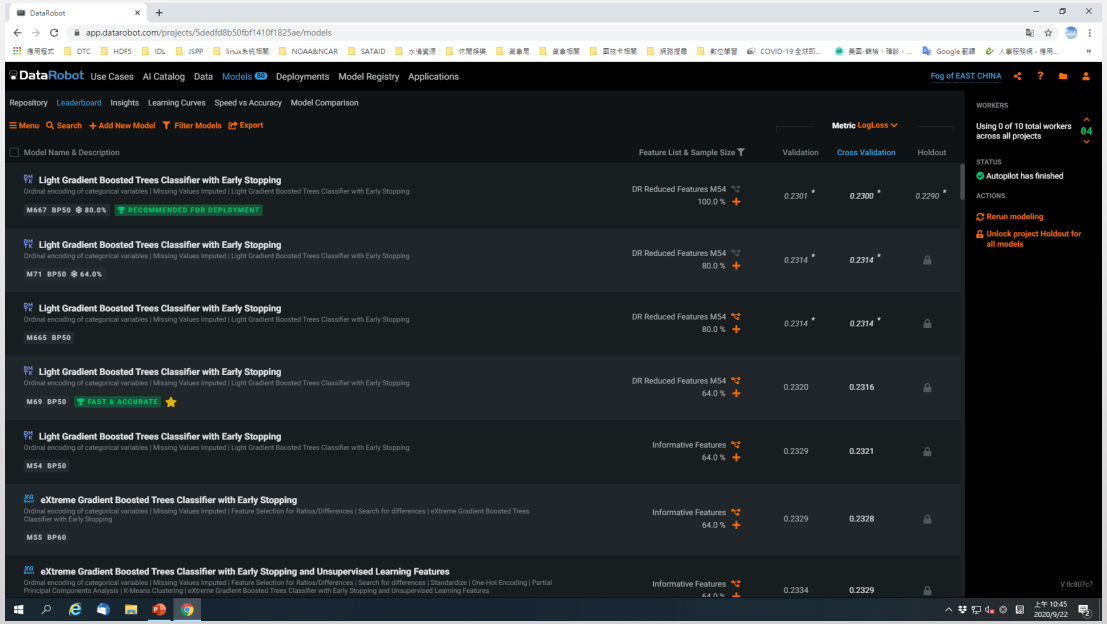
- 可自動執行模型建構過程
- 快速輕鬆地建構高準確度的模型

DataRobot的操作介面



資料上傳的畫面

只要將資料理成CSV或DataRobot認得的格式，透過滑鼠拖拉到頁面即上傳。



訓練結果的列表

列出所有演算法的結果並排序，效果最佳的排在上面。

DataRobot的學習結果

霧

說明：DataRobot的實驗以Features的內容分為Run0,Run1,Run2三個，右圖表列出經過訓練之後排名前面的三名的Model（以RMSE作為排序的依據）。

訓練資料

資訊時間：2019年1月到6月

資料範圍：東亞範圍

資料內容：氣象站觀測資訊（有霧及沒霧），氣象站所在位置之地理資訊、衛星觀測資料及Level2產品。

Feature List

Run0：所有資訊（共160項）

Run1：去除幾何資訊（去除方位角、天頂角、高度等，共保留155項）

Run2：去除幾何資訊及Level 2 產品（共保留83項）

Feature List name		Run0			Run1			Run2		
Features		160			155			83		
Best Model		Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2011	0.2012	0.2017	0.2034	0.2038	0.2042	0.2149	0.2150	0.2158
	AUC	0.9513	0.9513	0.9514	0.948	0.948	0.9484	0.9304	0.9309	0.9301
Second Best Model		Light Gradient Boosted Trees Classifier with Early Stopping			Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2013	0.2012	0.2019	0.2037	0.2042	0.2045	0.2149	0.2149	0.2159
	AUC	0.9518	0.9511	0.9520	0.9446	0.9469	0.9472	0.9289	0.9295	0.9291
Third Best Model		eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2014	0.2016	0.2022	0.2038	0.2036	0.2048	0.2150	0.2151	0.2157
	AUC	0.9501	0.9512	0.9503	0.9478	0.9483	0.9481	0.9305	0.9304	0.9306

常用的指標

Accuracy = (TP+TN) / (TP+FP+FN+TN)

最常用的指標, 但在某種類的資料很少時(資料不平衡時)會失效。

Precision = TP / (TP+FP)

機器學習判定有霧的樣本中實際有霧的比例。

Recall (Sensitivity) = TP / (TP+FN)

實際有霧的樣本中機器學習判斷正確的比例。

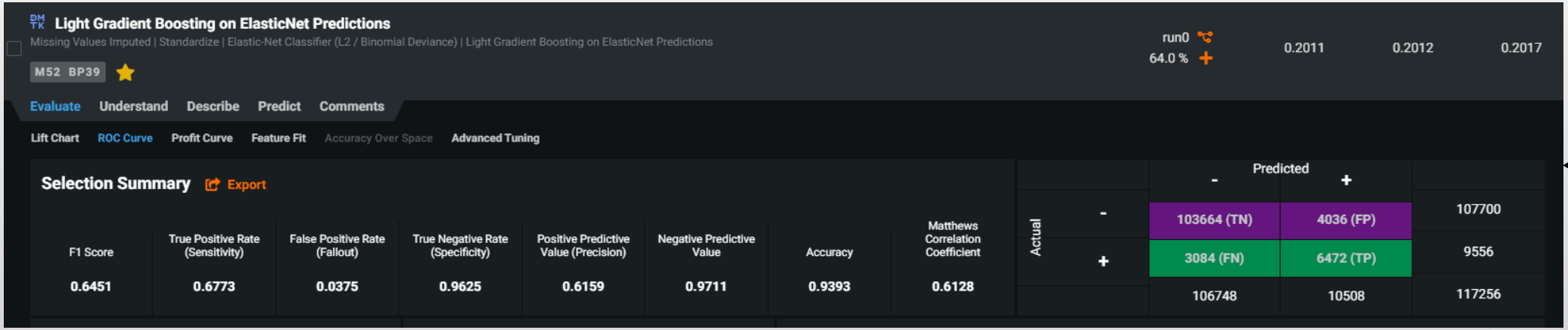
F1 Score $= \frac{2}{Recall^{-1}+Precision^{-1}} = \frac{2(Recall \times Precision)}{Recall+Precision}$

在資料不平衡的狀況下, 可參考F1 Score的表現, 是一種融合Precision及Recall的綜合指標。

		Predicted		
		-	+	TOTAL
Actual	-	**TN**	FP	TN+FP
	+	FN	**TP**	FN+TP
	TOTAL	**TN+FN**	**FP+TP**	

TN 機器學習和觀測均無霧。
TP 機器學習和觀測均有霧。

DataRobot介面提供的分析數據



DataRobot的學習結果

Feature List name	Run0	Run1	Run2
Features	160	155	83
Best Model	Light Gradient Boosting on ElasticNet Predictions Precision=0.6159 Recall=0.6773	Light Gradient Boosted Trees Classifier with Early Stopping Precision=0.6164 Recall=0.6535	eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) Precision=0.5543 Recall=0.6058
Second Best Model	Light Gradient Boosted Trees Classifier with Early Stopping Precision=0.6073 Recall=0.6891	Light Gradient Boosting on ElasticNet Predictions Precision=0.6088 Recall=0.65354	Light Gradient Boosted Trees Classifier with Early Stopping Precision=0.5389 Recall=0.6275
Third Best Model	eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features Precision=0.6258 Recall=0.6540	eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) Precision=0.599 Recall=0.6672	eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features Precision=0.5399 Recall=0.6302

Precision：

機器學習判斷有霧的樣本中觀測有霧的比例。

Recall (Sensitivity)：

觀測有霧的樣本中機器學習判斷正確的比例。

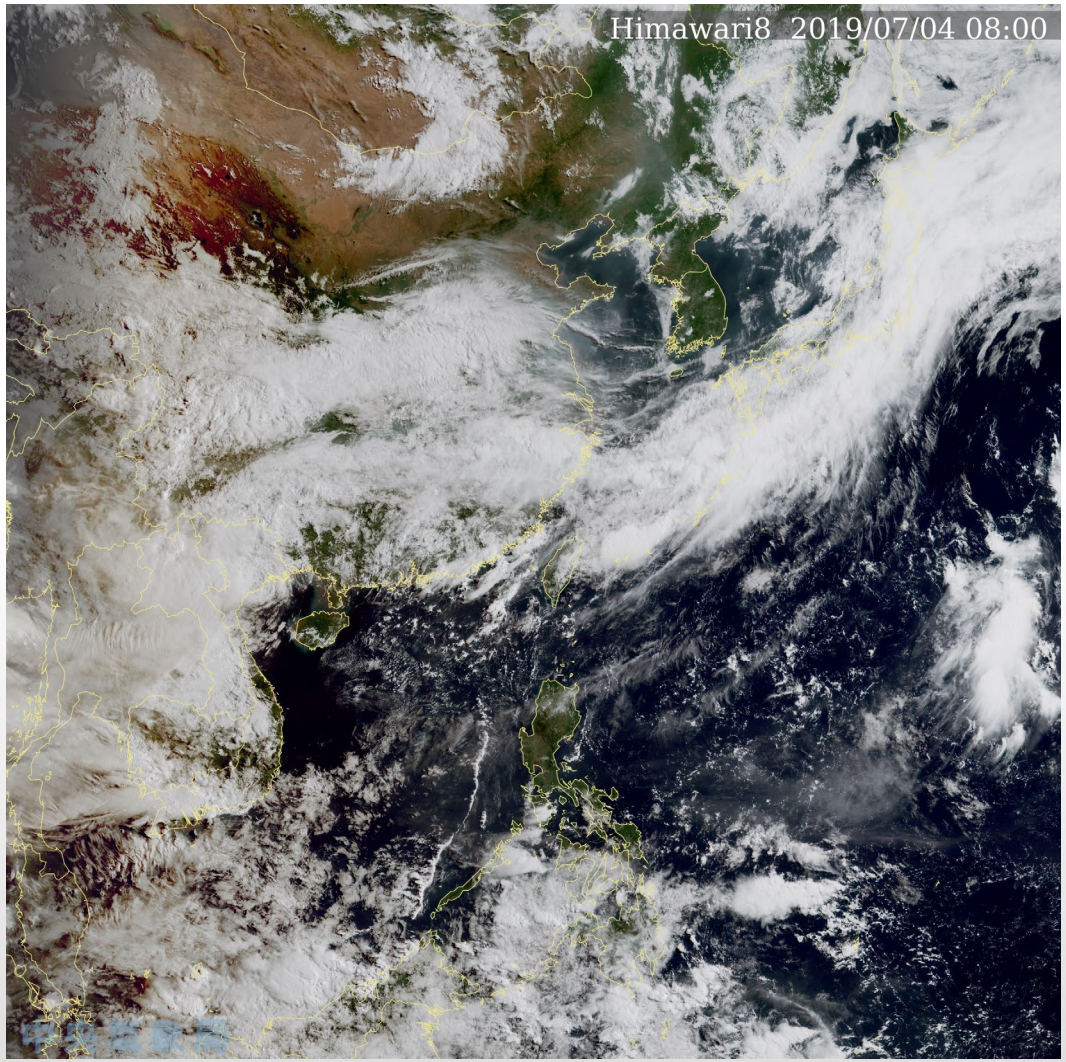
DataRobot的真實資料測試

測試資料

資訊時間：2019年7月4日

資料範圍：東亞範圍

資料內容：衛星觀測資料及Level2產品。



Run0：所有資訊（共160項）

最佳（RMSE = 0.2011）：
- Light Gradient Boosting on ElasticNet Predictions

次佳（RMSE = 0.2013）：
- Light Gradient Boosted Trees Classifier with Early Stopping

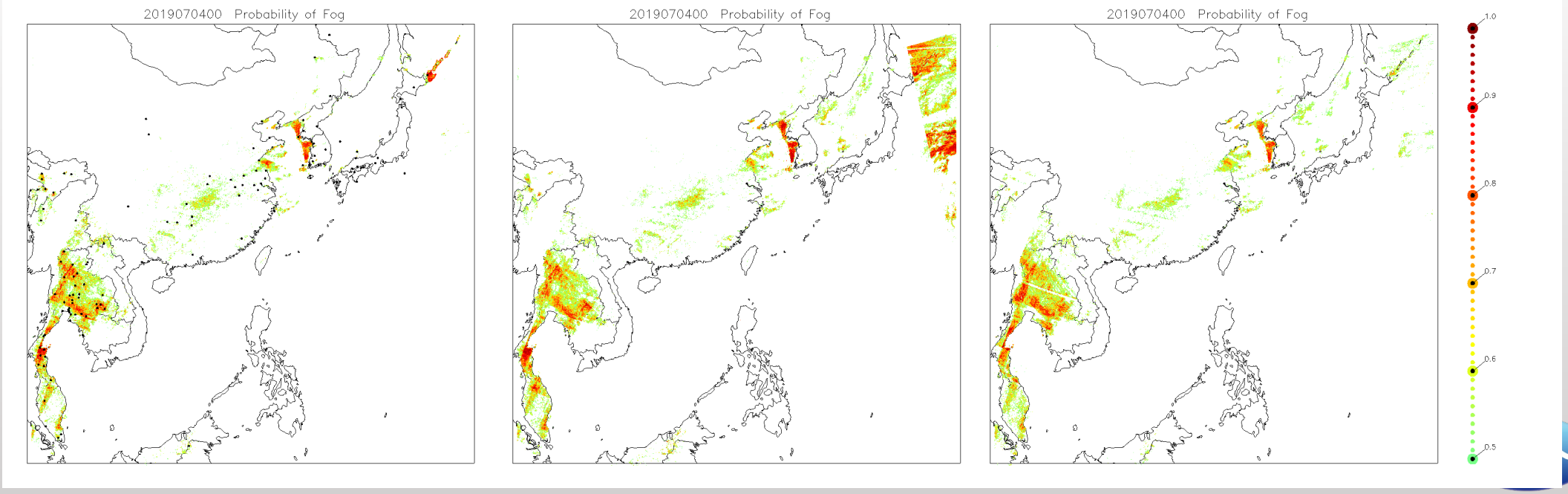
第三（RMSE = 0.2014）：
- eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features

Feature List name		Run0			Run1			Run2		
Features		160			155			83		
Best Model		Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2011	0.2012	0.2017	0.2034	0.2038	0.2042	0.2149	0.2150	0.2158
	AUC	0.9513	0.9513	0.9514	0.948	0.948	0.9484	0.9304	0.9309	0.9301
Second Best Model		Light Gradient Boosted Trees Classifier with Early Stopping			Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2013	0.2012	0.2019	0.2037	0.2042	0.2045	0.2149	0.2149	0.2159
	AUC	0.9518	0.9511	0.9520	0.9446	0.9469	0.9472	0.9289	0.9295	0.9291
Third Best Model		eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2014	0.2016	0.2022	0.2038	0.2036	0.2048	0.2150	0.2151	0.2157
	AUC	0.9501	0.9512	0.9503	0.9478	0.9483	0.9481	0.9305	0.9304	0.9306

Light Gradient Boosting on ElasticNet Predictions

Light Gradient Boosted Trees Classifier with Early Stopping

eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features



DataRobot的學習結果

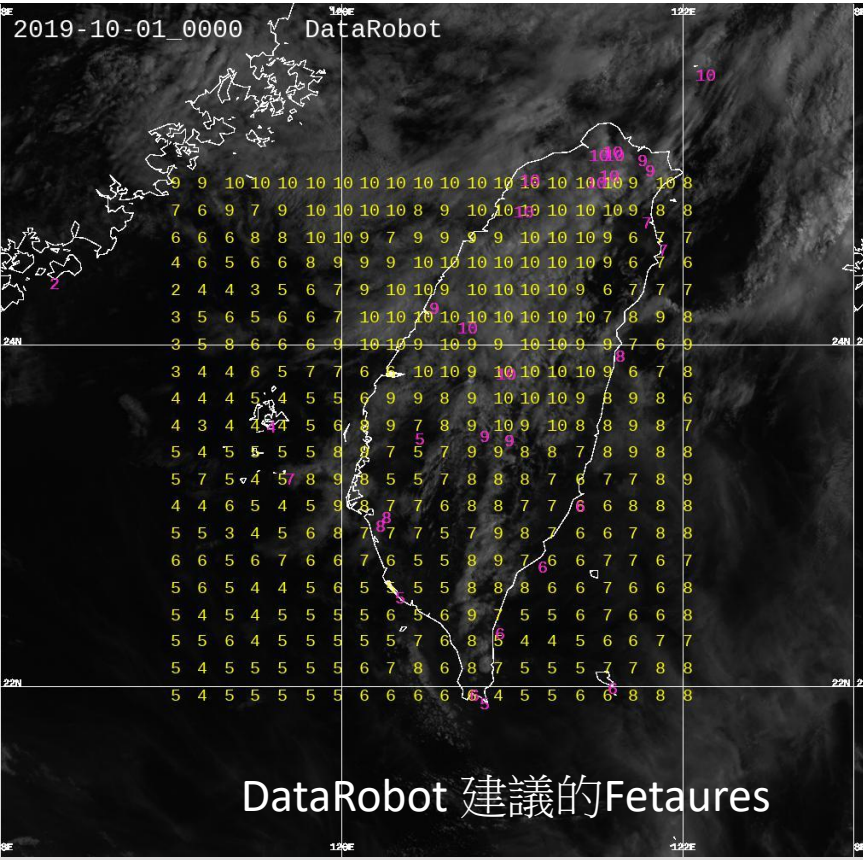
雲量

Fetures的組合	Features個數	驗證期間之RMSE		
		全日	白天	晚上
衛星中心目前提供的雲量資料		2.363	2.144	2.762
所有788個數據輸入	788	1.611	1.379	2.007
只有16個基本頻道	16	1.858	1.597	2.307
16基本頻道加上半徑4公里內的額外資訊	208	1.685	1.437	2.109
16頻道+半徑2公里資訊，無大值數量	48	1.85	1.553	2.351
4種基本頻道加上半徑4公里內的額外資訊	52	1.819	1.557	2.269
4基本頻道+2～12公里額外資訊	148	1.656	1.432	2.044
4基本頻道+2～14公里額外資訊	172	1.65	1.429	2.034
4基本頻道+2～16公里額外資訊	196	1.648	1.427	2.03
4基本頻道+2～16公里額外資訊+太陽天頂角	167	1.665	1.427	2.073
4基本頻道+10～16公里額外資訊	100	1.669	1.437	2.068
B07、B13、B15之 2～16公里平均值資訊	27	1.88	1.696	2.213
參考DataRobot建議名單，刪減部份特徵	96	1.621	1.383	2.026
參考DataRobot建議名單，刪減低貢獻特徵	88	1.634	1.387	2.056
參考DataRobot建議名單， 僅保留5種頻道（B03,B10,B11,B12,B15）	61	1.801	1.529	2.262
參考DataRobot建議名單， 僅保留6種頻道（B03,B07,B10,B11,B12,B15） 。	69	1.712	1.493	2.094
DataRobot建議名單（使用13種頻道）	101	1.613	1.375	2.019

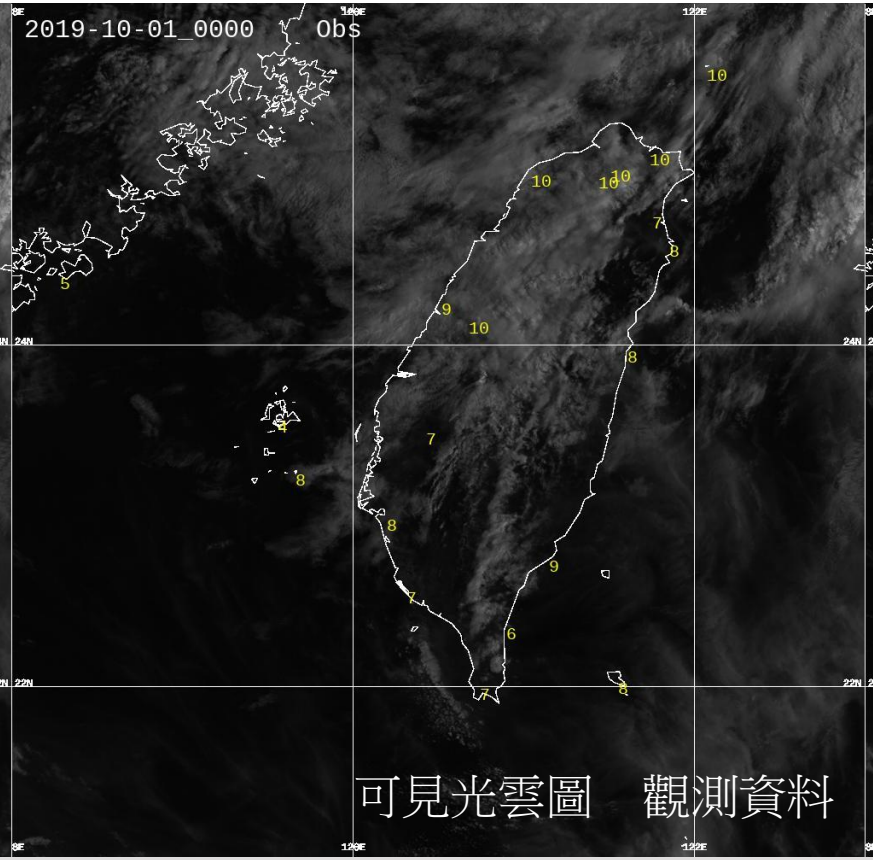
DataRobot的特色是能夠選出重要的Features。

DataRobot的學習結果

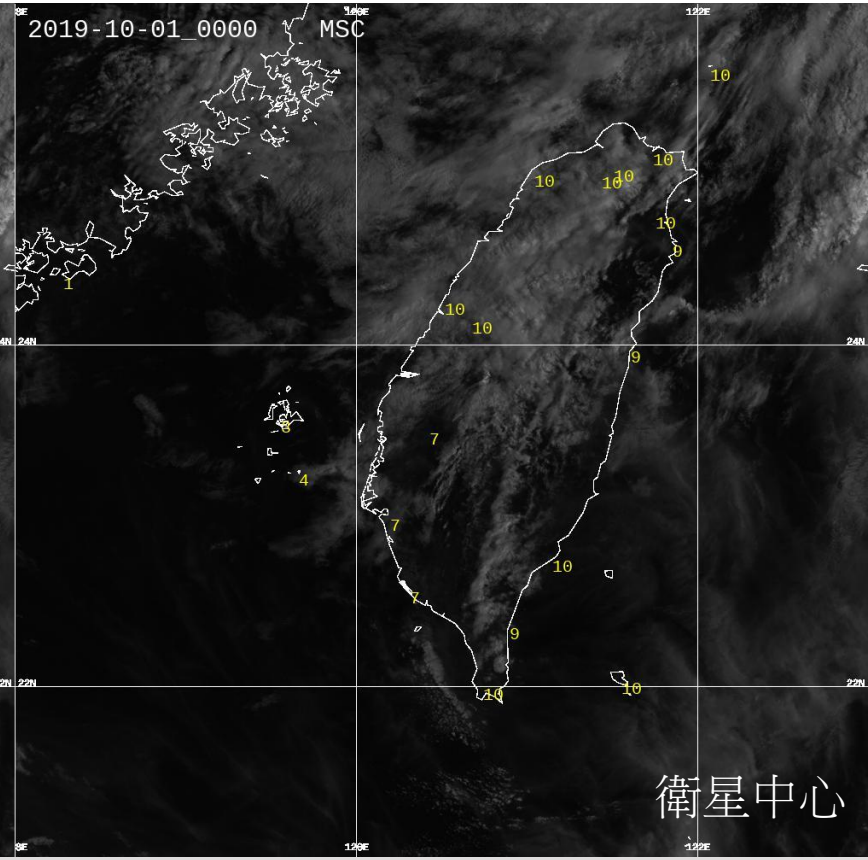
雲量



DataRobot 建議的Fetaures



可見光雲圖　觀測資料



衛星中心

結論

- DataRobot提供了一個很方便的平台。雖然有簡單的平台可用，但是仍需要處理大量訓練資料。
- 機器學習的效果比傳統的物理法稍好一點，是否值得取代物理方法的產品還需要審慎評估。
- 訓練須要有大量觀測資料做為訓練。

後續工作

- DataRobot的新功能增加影像的辨識。
- 衛星資料是數據也是影像。
  - 後續的思考方向為應用衛星的影象進行訓練。