2019
中華郵政
大數據競賽
分析方案說明書

郵件投遞是否有效率？
—降低郵件回投成本之策略初探

國立成功大學　來自很多系的隊

指導教授：鄭順林 老師

心理系 廖傑恩(隊長)／機械系 蔡詠丞／資訊系 張富嘉／統計系 郭士銘

目 錄

摘 要

郵務士投遞寄郵件時常發生無人接收而投遞失敗的問題，導致必須回投，其成本可觀。本研究旨在探究影響投遞成功與否的關鍵因素，並分成回投郵件與首投郵件來探討。

本研究使用中華郵政提供之郵件交寄資料(ACC)、特種郵件狀態查詢資料(TT)與自中央氣象局取得之公開的天氣資料(weather)及測站資料(stationInfo)，透過日期、時間、郵遞區號與郵件編號等資訊將上述資料串連與整理，將資料放入Random forest, Xgboost與LightGBM等機器學習模型中進行訓練，找出影響投遞成功與否的關鍵變項，再利用畫圖呈現出影響的機制。訓練出的模型也再透過model ensembing進行優化，提升預測表現。

本研究發現改變投遞時段與投遞星期可大幅降低投遞失敗率。以此發現，本研究提出一套「建議投遞時段系統」之理想系統的運作構想，期待能成為提升投遞成功率之智慧物流系統，促進郵務智慧化，甚至是自動化。

1 研究問題與提案動機

一、研究問題

郵件回投成功機率之探討：投遞時段、郵遞區號、天氣狀況與郵件種類等因子如何影響郵件是否成功，並分成回投郵件與首投郵件來探討。

二、提案動機與預期效果

郵務士(投遞員)寄送特種郵件時，時常發生無人接收而投遞失敗的問題，導致必須回投，即使每件郵件有三次的投遞數上限，但其成本依然可觀，這是中華郵政在「郵遞成本」與「郵遞效率」上難以突破的瓶頸。

雖預計民國108年底在全台設置超過2000座「i郵箱」，使物流配送的渠道多元化，但從 TT 資料集中的郵件狀態可知尚有大約15%投遞失敗的比例，故本研究想針對此問題進行探討，**利用投遞時段、地區（郵遞區號）、天氣、郵件種類等資料預測郵件「首投」與「回投」成功之機率。期望可優化郵遞流程、增加投遞員郵遞效率與降低郵遞成本，進而促使智慧物流之發展**。

＊註：若執行此方案勢必損害部分客戶之權益，故可能需搭配其他方案(Ex: i郵箱)一同配合，在「郵遞成本」與「客戶權益」間取得平衡，故此問題值得另外探討。

2 文獻回顧

中華郵政未來發展三大方向(2017/10/31)

中華郵政公司宣布，將積極推動業務轉型，朝「智慧物流」、「數位金融」及「長照服務」3大經營目標邁進。為順應社會新需求須轉型，在「智慧物流」方面，現正積極於全國佈建iBox(智慧取件箱)，預計民國108年底，將建置超過2000座，提供客戶多元選擇。

中華郵政董事長魏健宏四項發展重點(2018/5/11)

中華郵政公司董事長魏健宏推動郵務轉型，發展智慧物流業務，其最基本為民服務業務就是信件、包裹遞送，為讓遞送服務更有效達成、提高可靠度、降低成本、確保人員運輸安全，未來將參考智慧物流體系的架構與優點，做為改善郵務作業參考。

A Study on the Number of Domestic Food Delivery Services

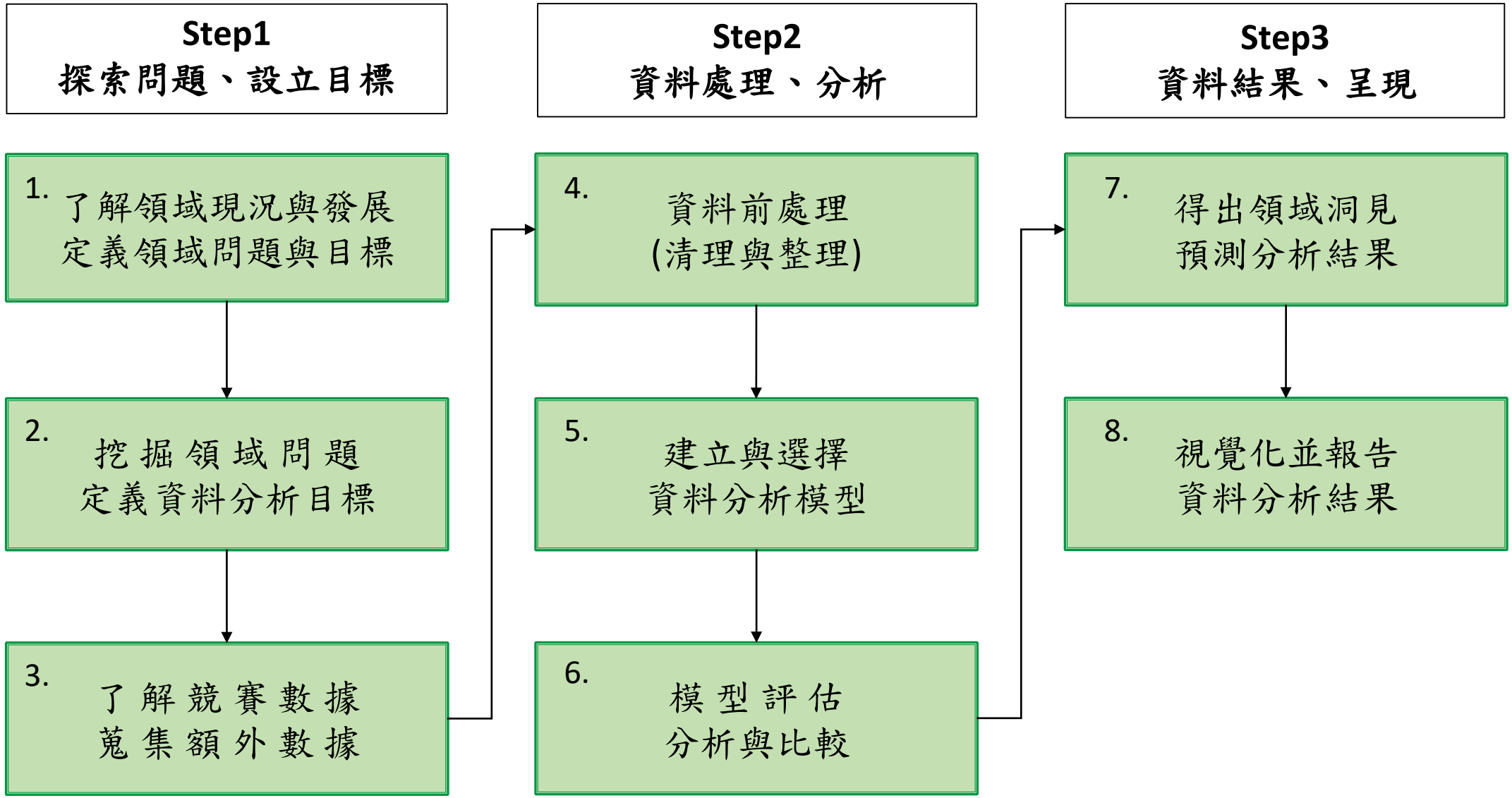
研究論文，關於韓國食品運送需求預測，與本提案類似，運用日期時間與天氣資料建立預測模型，使用的模型架構有：線性回歸、隨機森林、Gradient boosting、支援向量機（SVM）、類神經網路、Logistic regression。

文獻中顯示，他們建立在日期與時間因素的模型可以預測需求。

A Practical Guide to Support Vector Classification

一份關於支援向量機的實務使用教學，裡面包含了從資料前處理、模型的理論知識、重要參數、示範程式...等等的流程解說，還有實際案例供研究參考。

3 資料分析流程架構—分析流程圖



3 資料分析架構（1）

1. 了解領域現況與發展 定義領域問題與目標

- **中華郵政現況與發展：**

中華郵政以「智慧物流」、「數位金融」及「長照服務」三大經營為目標，積極推動郵務轉型，發展智慧物流業務。(中華郵政未來發展三大方向(2017/10/31))

- **中華郵政問題與目標：**

近年隨物聯網興起，電子商務蓬勃發展，帶動包裹遞送需求成長，欲透過資訊、通訊、大數據分析等技術，優化郵遞流程。讓「遞送服務更有效達成」、提高可靠度、「降低成本」、確保人員運輸安全，未來將參考智慧物流體系的架構與優點，做為改善郵務作業參考。(中華郵政董事長魏健宏四項發展重點(2018/5/11))

2. 挖掘領域問題 定義資料分析目標

- **挖掘郵遞問題與定義分析目標：**

藉由「智慧物流」替企業「提高效率」、「降低成本」，是物流業當前著眼的目標，其中值得關注的環節為「倉儲中心管理」與「郵物運輸配送」。(王繼祥(2015年5月).「物聯網發展推動中國智慧物流變革」研討會.中國物流技術協會華夏物聯網.pdf)

當今在郵遞階段(後者)之機械化流程與人力尚未完全緊密結合，而因應郵政本業及網購需求增加，未來郵遞環節也不可避免，故以此切入大會提供四份資料檔朝向「提高效率」、「降低成本」、「客製化」及「客戶滿意度」為方向探討，**希望透過資料分析優化「郵遞流程」、降低「郵遞成本」，促進智慧物流之發展。**

3. 了解競賽數據 蒐集額外數據

- **競賽數據資料：**

共有4份資料檔(ACC, GPS, TT, CC)，資料時間為107/1/1至107/3/31。

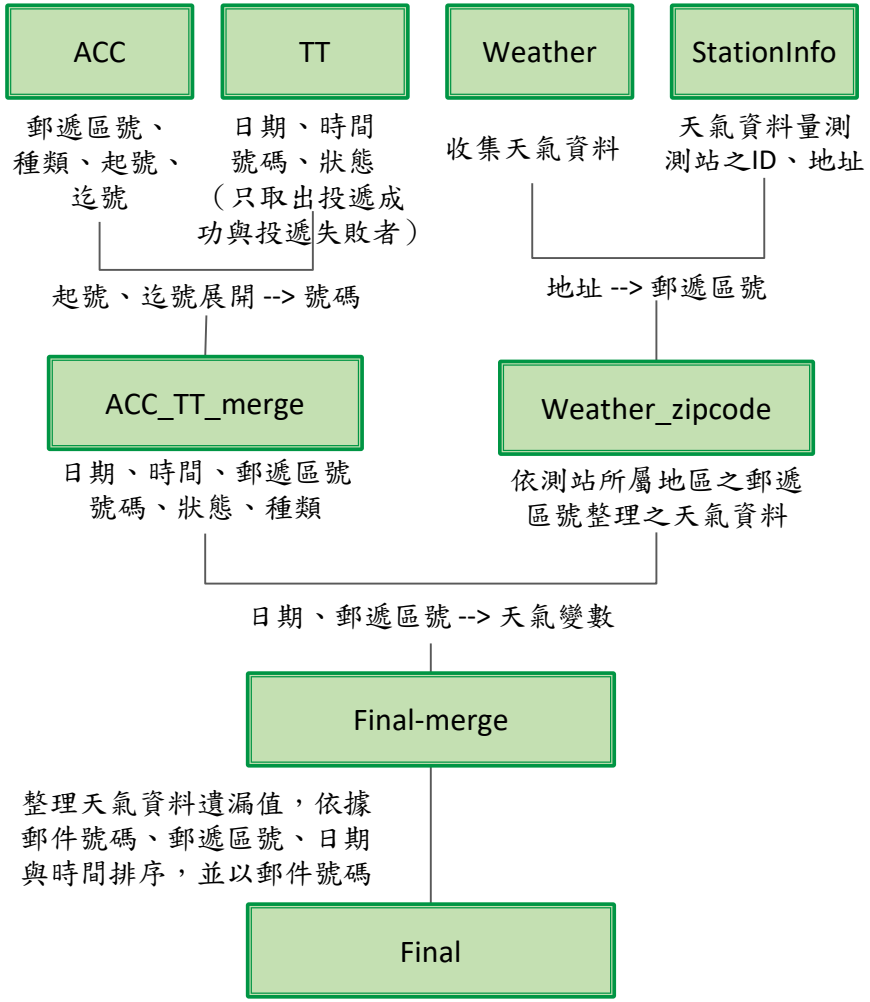
「郵物旅程」與「資料涵蓋範圍」：

- **額外數據資料：**

為增加預測的準確度及提高模型解釋的合理性，我們額外蒐尋「各地區天氣資料」及「天氣觀測站之地址」加以輔助。

3 資料分析架構（2）

4. 資料前處理 (清理與整理)



5. 建立與選擇資料分析模型

- 建模前之資料處理：

Scale

- 候選模型：

Logistic Regression	RandomForest	Catboost
LightGBM	XGboost	

1. 找出重要變項
2. 提升模型預測力

6. 模型評估分析與比較

- 模型評估標準：

Accuracy	Precision
f1Score	Cross-validaition

3 選用理論架構模型

選用理論架構模型

Logistic Regression	邏輯斯迴歸應用於二元分類的需求，能幫助預測投遞成功與失敗的分類，並計算出投遞成功的機率。
Support Vector Machine	SVM在工業上時常應用於分類問題，將原始資料(低維)投射到高維空間中，能更方便將各資料點的類別區分出來。
Boosting	Boosting 為機器學習中很重要的方法，其透過生成及組合多個較差分類器，使分類模型不斷進化，最後變成一個強大的分類模型。根據組合的方法、決策樹生成方式的差異，Boosting 相關的模型架構，都在考慮範圍內： 1. Random Forest 2. XGboost 3. LightGBM

處理環境（1）

- weather.csv資料處理：

Windows 10	RStudio(v1.1.456)	R(v3.5.1)
Anaconda1.9.7	Python3.6	Microsoft Office 365

- stationInfo & weather資料整合：

macOS(v10.14.3)	RStudio(v1.2.1511)	R(v3.6.0)
Microsoft Office 365		

- ACC.csv、TT.csv資料清理：

Windows 10	Anaconda1.9.7	Python3.6

- 最終資料(Final)整理與模型配適：

CentOS release 6.7	Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	Ram 288GB
R(v3.5.2)	Python3.4	

處理環境（2）

- R & Python Package:

（i）Python

ipython(7.3.0)	ightgbm(2.2.3)	matplotlib(3.0.3)
numpy(1.16.2)	pandas(0.24.2)	scikit-learn(0.20.3)
scipy(1.2.1)	seaborn(0.9.0)	statsmodels(0.9.0)
xgboost(0.82)		

（ii）R

dplyr(0.8.0.1)	data.table(1.12.2)	bit64(0.9-7)
ggplot2(3.1.1)	reticulate(1.12)	chron(2.3-53)
corrplot(0.84)		

4 資料描述與探索（1）競賽資料與額外資料

1. 競賽數據（大會提供）

- 共有4份資料檔：CC、ACC、GPS、TT
- 資料時間：107/1/1~107/3/31
- 資料筆數與變項數如下

資料	筆數 (#rows)	變項數	.csv資料大小
CC	46,325	28	5MB
ACC	27,979,054	38	4.13GB
GPS	48,861,658	16	5.55GB
TT	381,043,733	6	35.35GB

更詳細之大會資料探索請見資料描述與探索(3)

2. 額外數據：天氣資料

- 資料檔：weather
- 資料時間：107/1/1~107/3/31
- 資料筆數與變項如下

變項名稱	變項描述	型態
date	日期	date
hPa	氣壓(百帕)	float
Temper_C	氣溫(攝氏)	float
Wind	風量	float
Rain_mm	降雨量(毫米)	char
zipcode	測站郵遞區號	char

3. 額外數據：測站資料

- 資料檔：stationInfo
- 資料時間：107/1/1~107/3/31
- 資料筆數與變項如下

變項名稱	變項描述	型態
站號	觀測站之站號	char
站名	觀測站之站名	char
海拔高度	(單位公尺)	char
經度	12x.xxx	float
緯度	2x.xxx	float
城市	所在城市	char
地址	所在地址	char

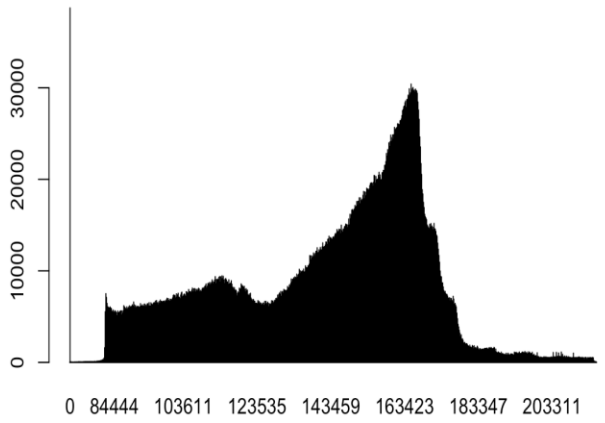
4 資料描述與探索（2）大會資料基本探索

更詳細之大會資料基本探索請點此連結： https://pse.is/HNXCM

NCKU 中華郵政

ACC.csv

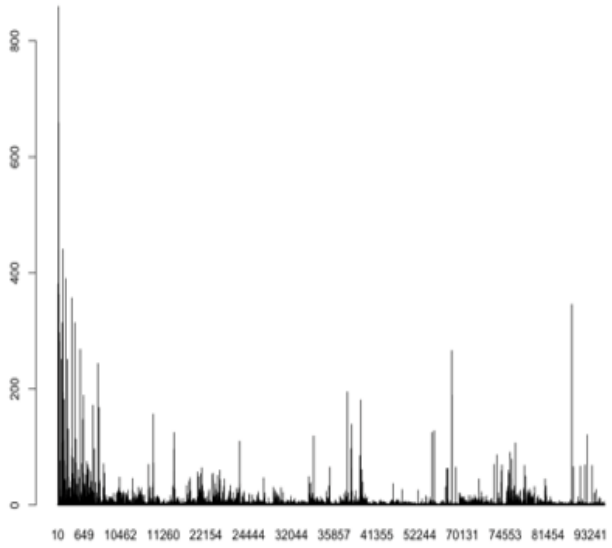
交寄時間之Barplot



交寄時間之平均數為 14：16：24，而大部分的交寄時間落在16~17點，之後則驟降，由上圖可知晚上交寄郵件的人較少。

CC.csv

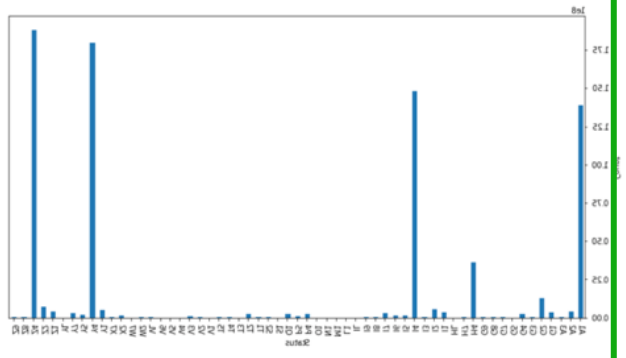
客戶郵遞區號之Barplot



應為3碼或5碼，之後需要處理碼數不對的。長條圖顯示有幾個特別大宗的郵遞區號，反應出這些郵遞區號代表的地區特別多特約客戶，其中「104」為眾數。

TT.csv

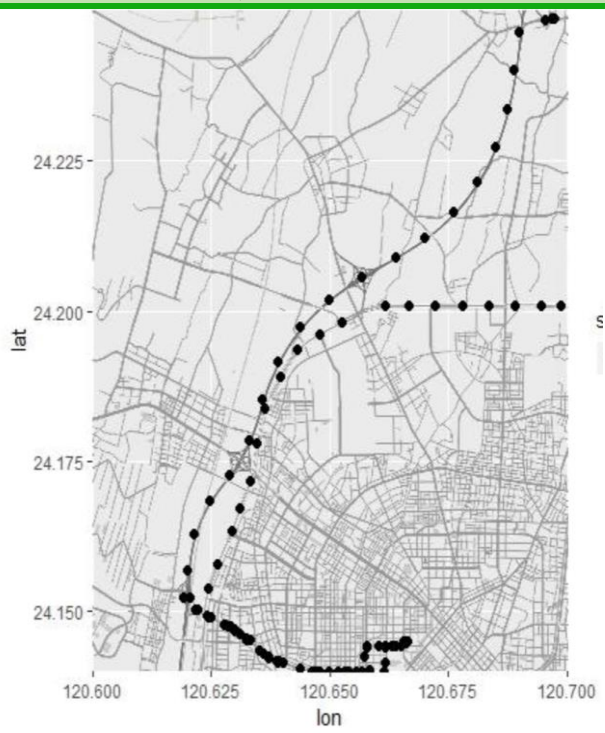
郵件狀態碼之Barplot



根據郵件狀態碼進行數量統計，統計結果如下長條圖所示，由長條圖可以看出：A1、H4、I4、Y4、Z4等狀態分別要較高的數量。(郵件狀態碼可參考中華郵政大數據競賽資料欄位規格之說明)

GPS.csv

GPS車輛軌跡圖



此圖示例：一輛四輪車輛於2018年一月一日行駛於道路上之衛星軌跡圖。

4 資料描述與探索（3）

4. 整理合併後之最終數據

- 資料檔：Final
- 資料時間：107/1/1~107/3/31
- 資料筆數與變項數：

  Final 有14,006,444筆、15個變項。

5. 因時間與計算成本限制，為簡便分析，初賽先進行台北市的資料的分析。以「郵遞區號<200」取出台北市的資料，數據狀態：

- 資料檔：Final_TP
- 資料時間：107/1/1~107/3/31
- 資料筆數與變項數：

  Final_TP有4,274,983筆、15個變項。

因有許多郵件號碼重複，但其對應到的郵遞區號又有差異，應為不同郵件，故將郵遞區號後方加上5碼郵遞區號作為新的郵件號碼，以更方便區別個別郵件。

各測站天氣資料與郵件查詢資料之合併係透過各測站地址所屬之5碼郵遞區號與日期。透過這些5碼郵遞區號在各日期的天氣資料可計算出各3碼郵遞區號（也就是各鄉鎮市區）在各日期的天氣資料平均值，無測站的5碼郵遞區號之天氣資料即取用這些平均值來代表。

是否為回投郵件係透過相同郵件號碼的日期差來決定，並考慮休假日，若工作日天數差小於等於一天，時間較晚的那一列稱為回投郵件。

變項名稱	變項描述	型態
zipcode	郵遞區號(五碼)	factor
date	日期	Y-m-d
Mail_No_zipcode	加上5碼郵遞區號之郵件號碼	factor
statusNum	郵件投遞是否成功	binary
time	時間	h:m:s
cate	基本郵件種類與細分類之合併	factor
Double	雙掛號類型（回執信件）	factor
Speed	是否為限時郵件	binary
hPa	氣壓(百帕)	float
Temper_C	氣溫(攝氏)	float
Wind	風量	float
Rain_mm	降雨量(毫米)	float
zipcode_3D	郵遞區號(三碼)	factor
time_group	將時間以四分位數切為四段	factor
Redelivery	是否為回投郵件	Binary

4 資料描述與探索（4）

Final_TP整體資料基本描述

Observed variables	Mean/percentage (SD)	No. of NA
Overall (n=4,274,983)		
statusNum	0.86 (0.35)	0
Double	0: 0.9802; 1: 0.0181; 2: 0.0013; 3: 0.0002; 4: 0.0003	0
Speed	0.13 (0.34)	0
hPa	1013.94 (7.41)	941,770
Temper_C	17.36 (4.09)	941,770
Wind	1.95 (1.04)	941,770
Rain_mm	4.41 (11.62)	475,940
Redelivery	0.13 (0.33)	0

Final_TP回投郵件資料(Redelivery=1)基本描述

Observed variables	Mean/percentage (SD)	No. of NA
Redelivery = 1 (n=550,670)		
statusNum	0.52 (0.5)	0
Double	0: 0.9813; 1: 0.0173; 2: 0.0012; 3: 0.0001; 4: 0.0001	0
Speed	0.11 (0.31)	0
hPa	1013.58 (9.11)	130,094
Temper_C	17.13 (4.02)	130,094
Wind	1.95 (1.01)	130,123
Rain_mm	4.2 (10.43)	65,687
Redelivery	1 (0)	0

Final_TP首投郵件資料(Redelivery=0)基本描述

Observed variables	Mean/percentage (SD)	No. of NA
Redelivery = 0 (n=3,724,313)		
statusNum	0.91 (0.29)	0
Double	0: 0.9799; 1: 0.0182; 2: 0.0014; 3: 0.0002; 4: 0.0003	0
Speed	0.13 (0.31)	0
hPa	1013.99 (7.13)	811,676
Temper_C	17.4 (4.09)	811,676
Wind	1.96 (1.04)	811,910
Rain_mm	4.44 (11.78)	410,253
Redelivery	0 (0)	0

1. 可以看出在台北市的投遞資料中，回投郵件投遞成功率比首投郵件投遞成功率低許多。
2. 天氣資料雖然已透過取較大地區之平均來降低遺漏值數量，但由上三表可知天氣資料之遺漏值數量仍多，故分析上宜採用較不受遺漏值影響之tree與boosting的分析方式。

4 資料描述與探索（5）

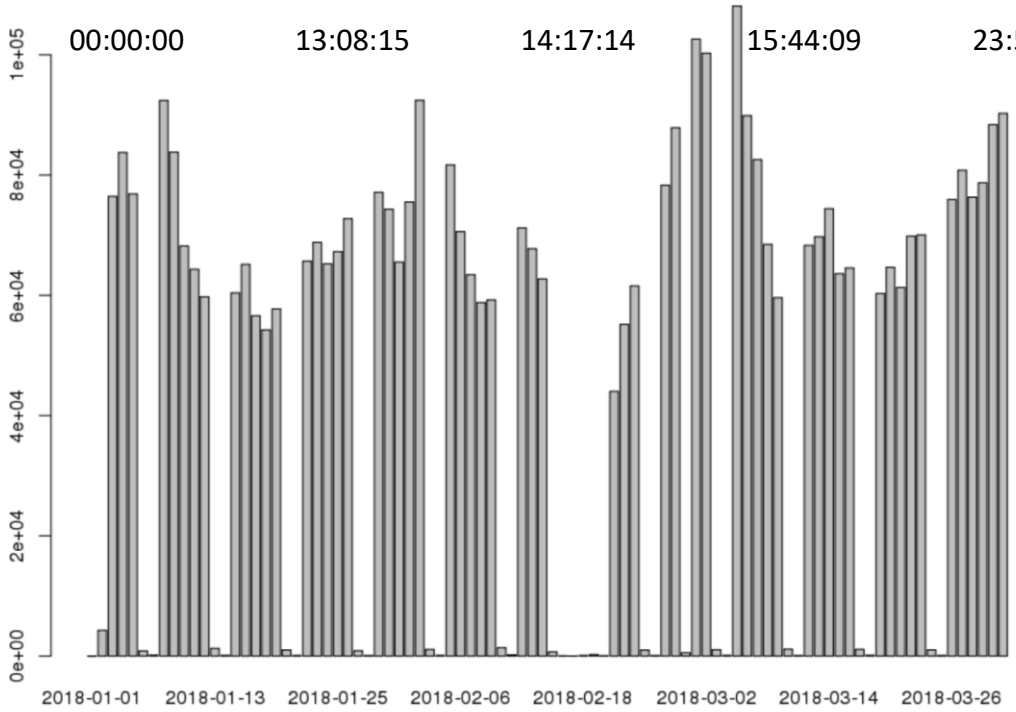
Final_TP 整體資料日期與時間的基本描述

日期 (年-月-日)

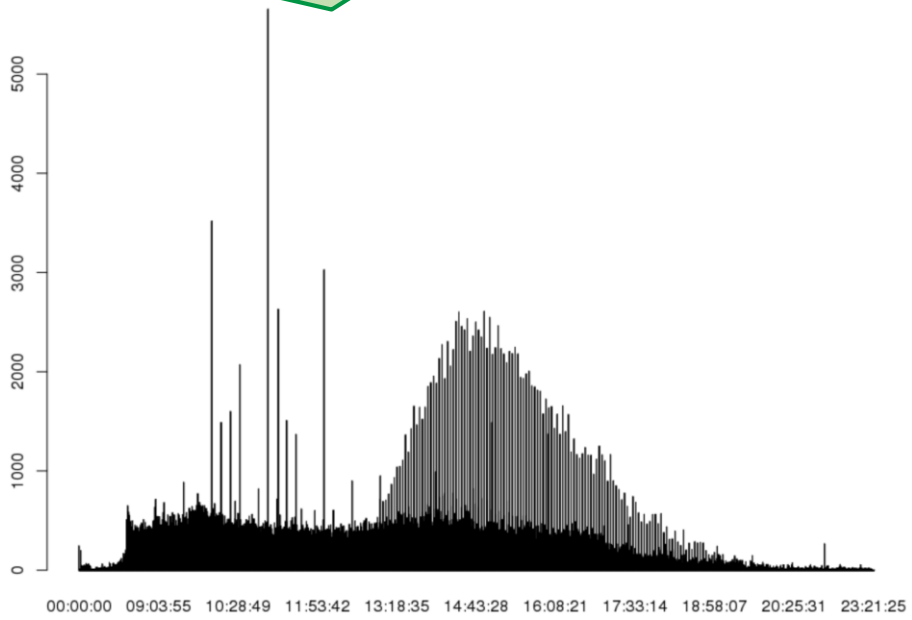
Min (始)	第一分位數	中位數	第三分位數	Max (終)
2018-01-01	2018-01-24	2018-02-14	2018-03-13	2018-03-31

時間 (時:分:秒)

Min (始)	第一分位數	中位數	第三分位數	Max (終)
00:00:00	13:08:15	14:17:14	15:44:09	23:53:26



由下圖可看出郵件狀態時間的分布在上午有幾個尖峰，但整體資料的分布多集中在下午。

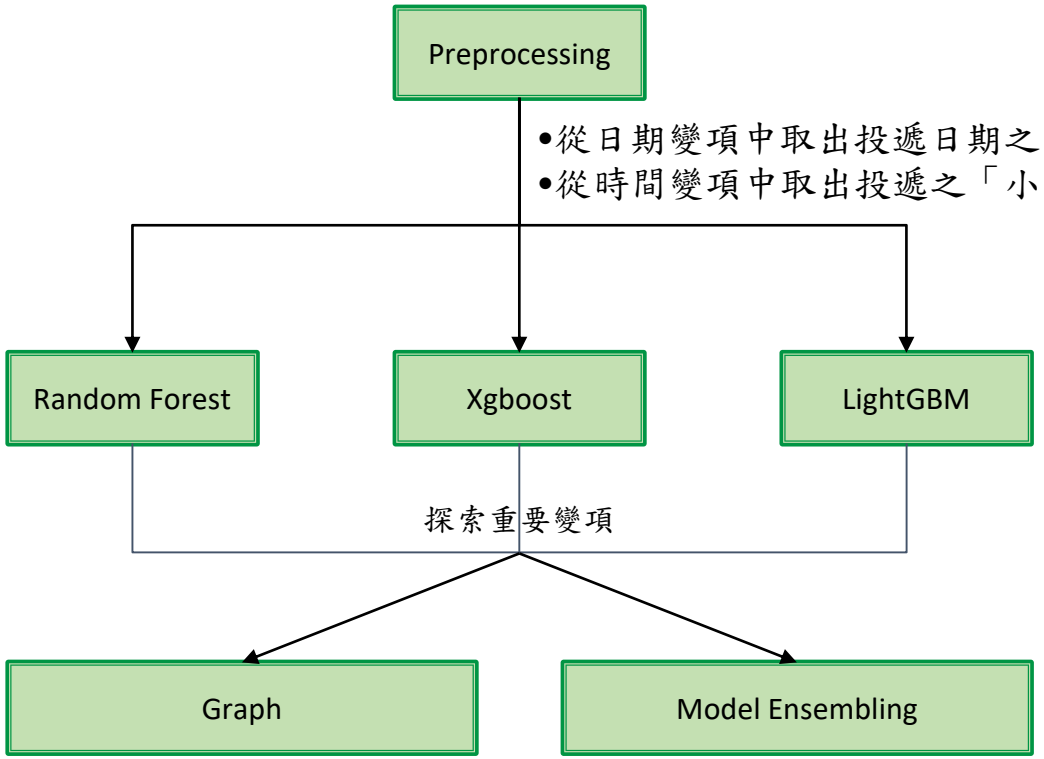


郵件狀態日期具有明顯的週期性（週末資料量少），而2月中旬由於農曆新年的關係有段資料空窗。

4 資料描述與探索（6）

中華郵政

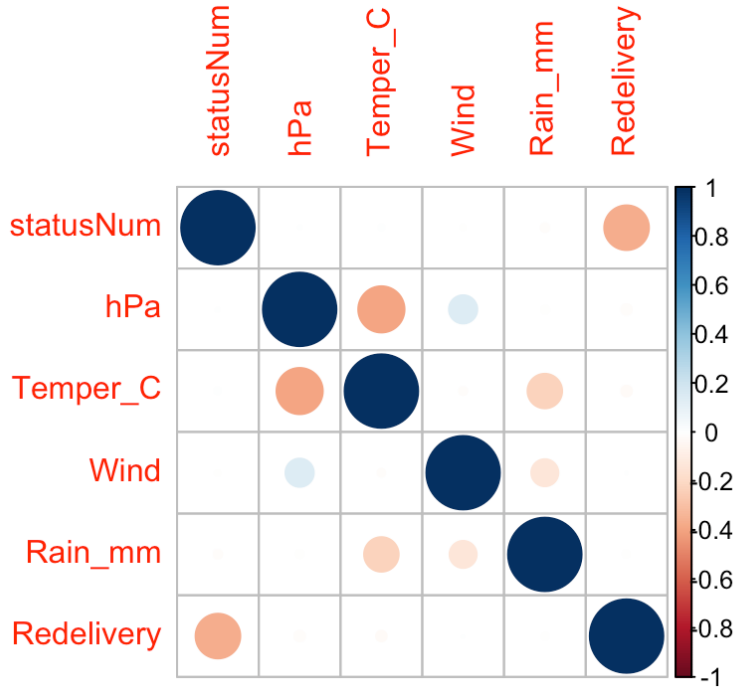
模型訓練流程



透過視覺化圖表呈現變項中影響之關鍵類別與影響程度

優化模型預測表現

視覺化圖表



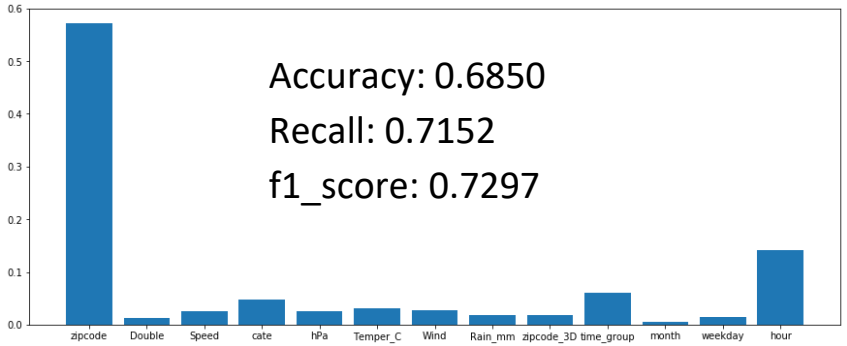
天氣資料之變項間無明顯高相關，無多重共線性之問題，故一同加入模型中。

5 資料分析結果與討論（1）

台北市整體資料（不分是否回投）

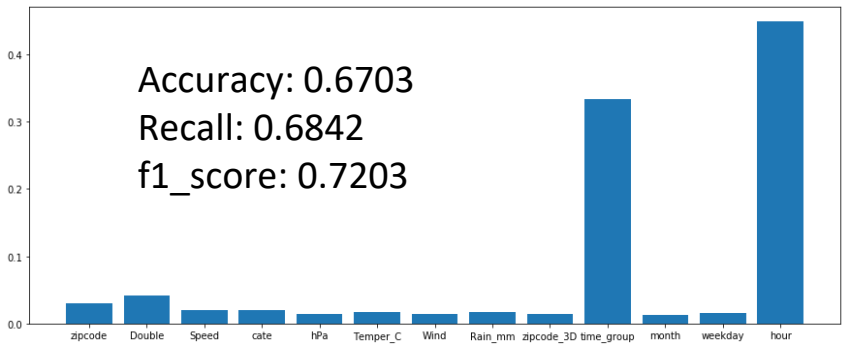
Random Forest 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	**82055**	**85015**
實際投遞成功	**306037**	**768422**



XGboost 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	**97129**	**69941**
實際投遞成功	**339351**	**735108**



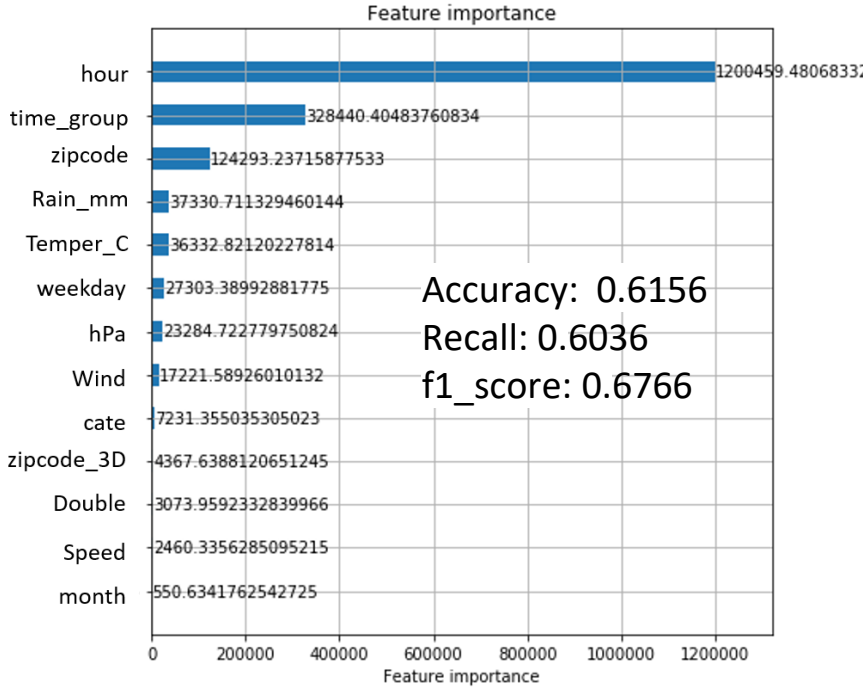
Model Ensembling 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	**67823**	**99247**
實際投遞成功	**207032**	**867427**

Accuracy: 0.6297
Recall: 0.7940
f1_score: 0.6161

LightGBM 結果：

	預測投遞失敗	預測投遞成功
實際騰地失敗	**115747**	**51323**
實際投遞成功	**425908**	**648551**

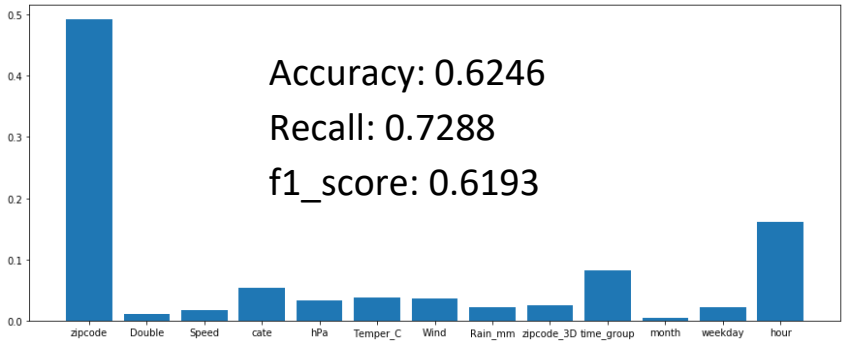


5 資料分析結果與討論（2）

台北市回投(Redelivery=1)郵件資料

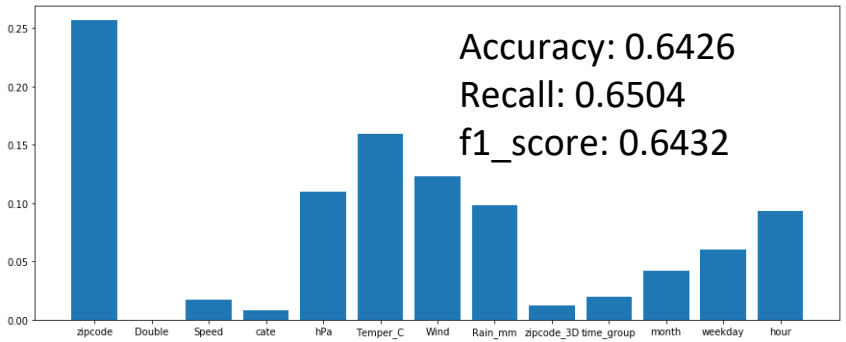
Random Forest 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	35478	35672
實際投遞成功	23345	62739



XGboost 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	45050	26100
實際投遞成功	30088	55996



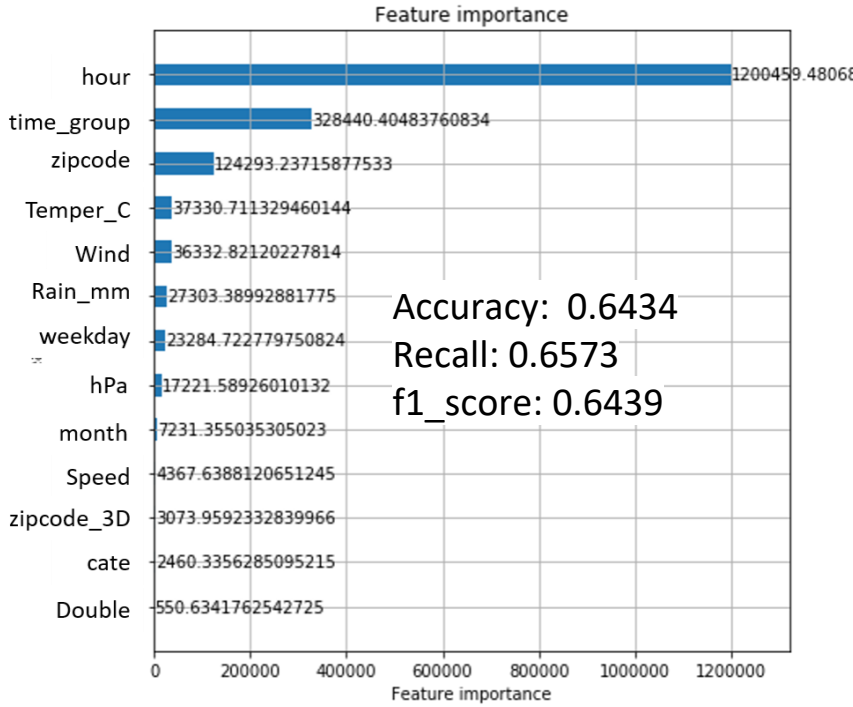
Model Ensembling 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	30668	40482
實際投遞成功	17731	68353

Accuracy: 0.6297
Recall: 0.7940
f1_score: 0.6161

LightGBM 結果：

	預測投遞失敗	預測投遞成功
實際騰地失敗	44577	26573
實際投遞成功	29494	56590

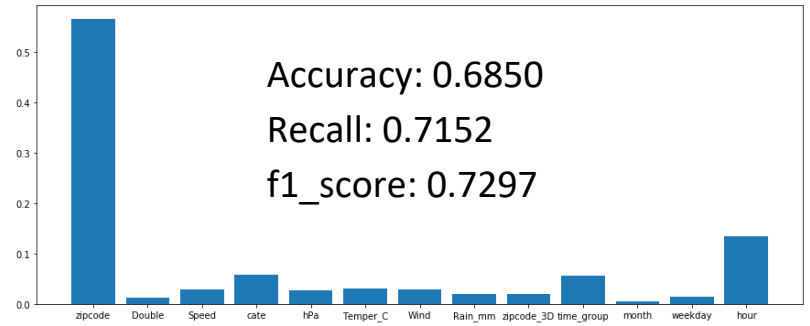


5 資料分析結果與討論（3）

台北市無回投(Redelivery=0)郵件資料

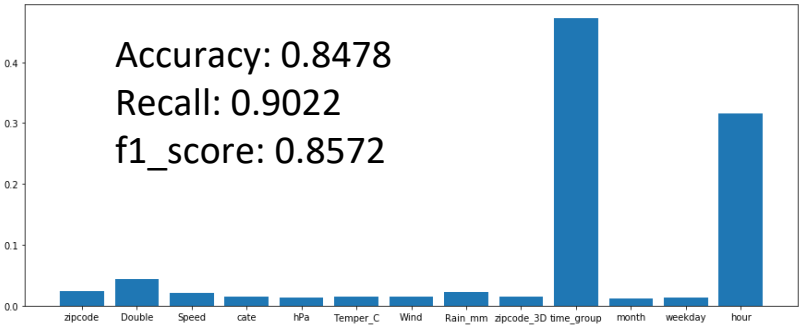
Random Forest 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	82055	85015
實際投遞成功	306037	768422



XGboost 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	23524	67877
實際投遞成功	97108	895786



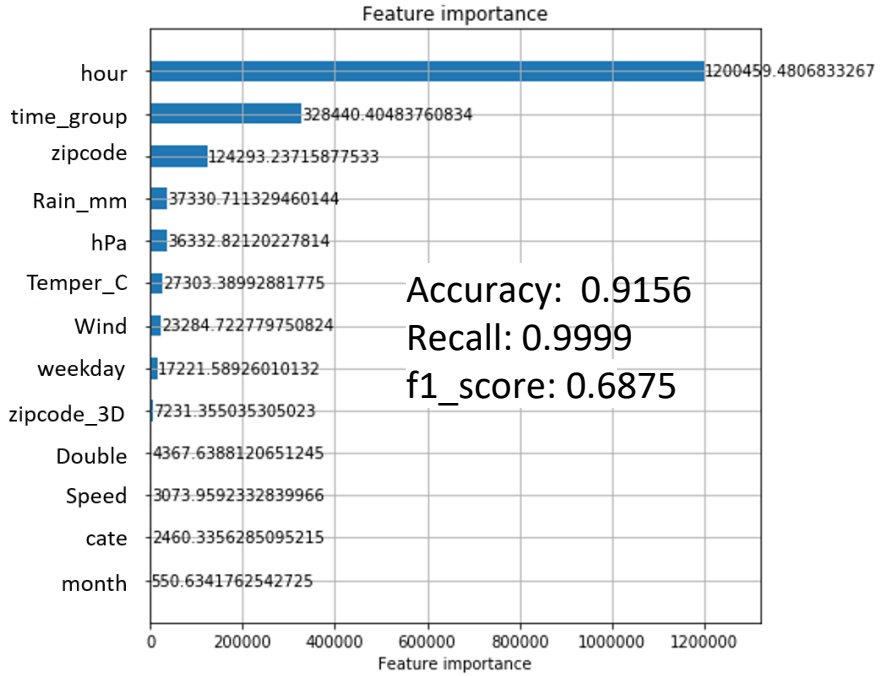
Model Ensembling 結果：

	預測投遞失敗	預測投遞成功
實際投遞失敗	38	91363
實際投遞成功	92	992802

Accuracy: 0.9156
Recall: 0.9999
f1_score: 0.8755

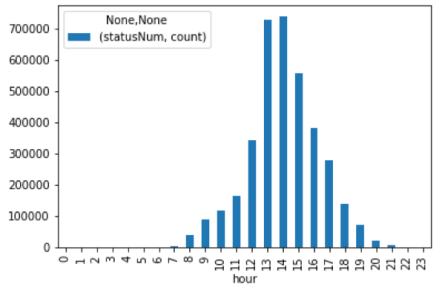
LightGBM 結果：

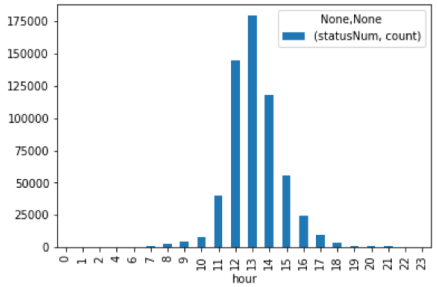
	預測投遞失敗	預測投遞成功
實際騰地失敗	40	91361
實際投遞成功	113	992781



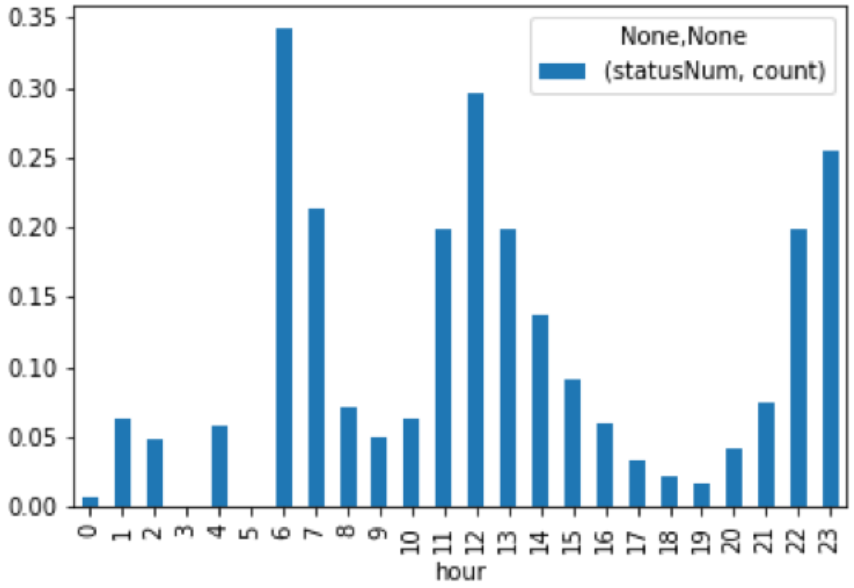
5 資料分析結果與討論（4）

透過模型訓練可以發現，在台北市的郵件中，不論是整體來看，或是分成回投郵件與首投郵件來看，郵遞區號、投遞時段與小時、投遞日的星期都有被認為是重要變項，而在回投郵件中，回投當日該郵遞區的氣壓、氣溫、風量與降水量也有被認為是重要變項。關於我們再透過畫圖來進一步探索與呈現。



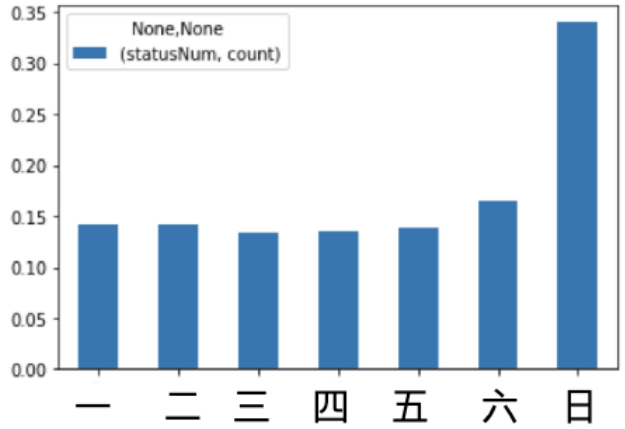


左上兩小圖為以小時為間格的郵件投遞成功(左)和失敗(右)數量長條圖，可以看到大部分的郵件寄送都集中在白天(8-19點)，且於中午時段達到高峰。



左下大圖則為以小時為間格的郵件投遞**失敗率**長條圖，可以看到在白天當中，中午(11~14點)的失敗率較高。

下圖為各投遞日期之星期的失敗率，可以看出星期日投遞之失敗率明顯高於其他星期，且超過2倍。

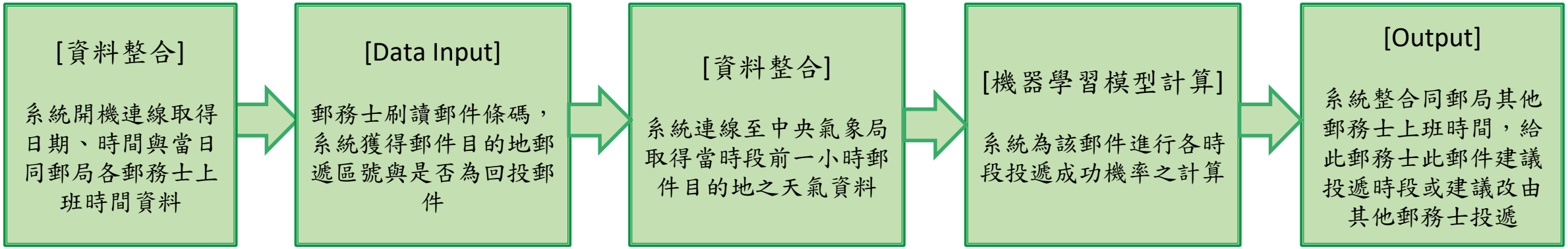


小結

中午(11-14時)是投遞尖峰的時段，然後此時投遞失敗率高，若能分配部分中午投遞的郵件改在早晨(8-10時)及下午(16-19時)投遞，除了能舒緩中午時段之投遞量使其不致超過負載量，也能大幅提升整體郵件成功投遞的次數。投遞日期的部分則建議減少在星期日投遞的數量以降低失敗率。

5 資料分析結果與討論（5）

本研究希望透過資料分析優化「郵遞流程」、降低「郵遞成本」，促進智慧物流之發展。透過分析，我們發現改變投遞「時段」與投遞「星期」應有助於郵件投遞之成功，進而大幅降低成本。經由此研究之模型，未來再優化其可行性與可靠性，並整合為「建議投遞時段系統」使郵務士使用之參考，進而促進智慧物流之發展。此系統理想運作流程如下圖所示。



在給定日期、時間、郵件目的地郵遞區號、是否為回投郵件，整合郵務士上班資料與天氣資料後，理想中的系統可計算出該郵件在各時段投遞成功機率，並依據該郵局各郵務士上班時段與負責郵件數量，給出投遞時段或改由其他郵務士投遞的之建議。**如此可大幅提高郵件投遞成功率，也考慮了郵務士們的工作量分配，是結合「優化郵件投遞流程」與「降低郵件投遞成本」的智慧物流系統。**

6 未來展望與心得回饋

未來展望	❑ 不僅限於台北市資料，而是針對全國資料的深度分析。 ❑ 嘗試更多模型參數的組合，並提高模型驗證次數，以提升模型表現。 ❑ 深入探討各郵遞區號與投遞時段間的交互作用。 ❑ 深入探討天氣變項對回投郵件投遞成功造成的效果
隊員們之心得	❑ 心理系 廖傑恩：人類行為乃至於社會上的各種運作背後都有神奇的故事，透過大數據，我們可以看到這個故事，而得到更多對未來有用的啟發。 ❑ 資工系 張富嘉：處理數據就和寫程式一樣，需要滿滿的熱情和耐心。 ❑ 機械系 蔡詠丞：咖哩的美味，在於香料組合間的調和與烹煮。資料也是，想法也是。 ❑ 統計系 郭士銘：利用統計專業所學，透過資料分析，讓數據說話！
針對資料收集與資料品質的建議	❑ GPS資料集中的狀態代碼，如果廠商或是駕駛願意回報狀況的話，這筆資料能透露出的訊息會增加非常地多，如果只是自動回報行駛或是暫停狀態，在資料的意義上，僅餘追蹤車輛之功能而已。建議可以訓練駕駛或與 GPS 系統商洽談。考慮到狀態分類較多，可針對幾項較為代表性之狀態碼進行回報。 ❑ GPS資料集中可紀錄車輛的用途，如郵務／公務。 ❑ CC資料內可向客戶調查滿意度，在合約之後的解約/續約原因中，在數據上可以計算出郵局之優劣與挖掘郵局的潛在客戶。 ❑ TT 資料集中，混進了非競賽標示時段(107/1/1-107/3/31)內的資料。 ❑ 掛號郵件條碼應統一由郵局處理，不應單單使用民眾自訂之條碼，否則會造成郵件難以被個別辨識，失去許多分析所需的資訊。建議可在民眾自訂碼後再加入流水號，方便辨識。