三個船長

團隊成員

李宜臻
張仲樸
潘拓雲
潘沐宣

大綱

- 背景簡介
  - 作品名稱
  - 提案動機
  - 摘要
- 資料分析流程圖
- 探索式資料分析(EDA)
  - 資料分析報表結果證明
- 模型建立
  - 處理環境
  - 資料清理及因子(feature)選擇過程說明
  - 模型選擇原因說明
  - 模型預測結果
- 結論

作品名稱及提案動機

作品名稱

透過資料探勘及數據模型，探討影響掛號郵件成功/失敗(成功率)的因素

提案動機

郵務運輸具有運量龐大(特種郵件追蹤查詢資料表中，一季的資料高達8900萬筆不同掛號號碼)、環節複雜(郵件狀態代碼表中共有63種不同狀態代碼)等特性。在郵務競賽的新聞稿中，郵政訓練所所長劉錫標提到希望能藉此競賽降低郵務成本，並提高準確率之目的

而我們在探索式資料分析的過程中，發現在一季的資料內，有高達12.06%的郵件曾經投遞失敗，此一投遞失敗的結果，不但影響投遞成功率，也增加投遞失敗後之人力及時間處理成本，影響層面甚大，是個重要且必須解決的問題。因此我們希望透過資料發現在郵務流程中存在的問題，找出影響郵件投遞成功/失敗(成功率)的因素，以發想最適當的解決方案，提升投遞成功率

Source: <Digitimes: 中華郵政70餘年數據 如何百鍊成金, 2019-03-07>

摘要

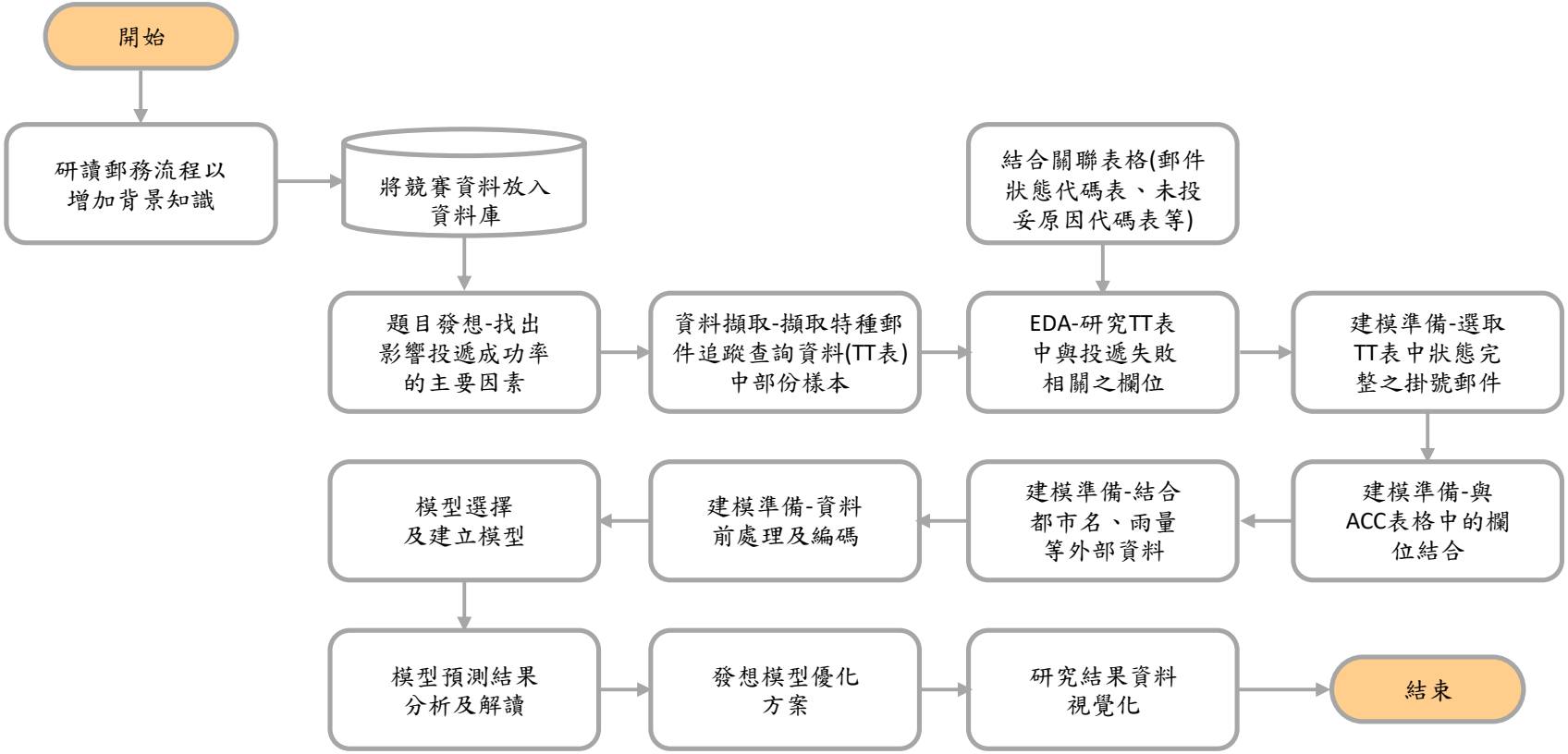
本研究主要探討掛號投遞過程中可能經過的各種環節及遇到的問題，找出影響郵件投遞成功/失敗(成功率)的因素。我們選擇特種郵件追蹤查詢資料表(TT表)作為研究對象。在TT表的母體中，我們計算各個郵件狀態代碼，了解掛號在投遞過程中最常出現的狀態分布後，發現曾經投遞失敗的掛號比例相當高達12.06%，因而選擇研究「投遞失敗」相關議題作為我們探討面向

因TT檔案大小高達34GB，受限於電腦算力，若處理整份檔案恐相當耗時，因此我們將TT檔案放置在Postgre SQL資料庫上，隨機抽樣進行探索式資料分析(EDA)及建立模型。我們在EDA時發現曾經投遞失敗的郵件

1. 使整體處理時間提高到5.9至數十天不等
2. 投遞失敗的原因有高達83%以收件人「不在」及「按鈴無回應」所造成
3. 投遞時間在中午11點-14點時段，投遞失敗的機率最高

為研究是否有其他因子對於投遞成功率造成影響，我們結合掛號之屬性、寄達地區及雨量資料，透過5種分類模型預測每封掛號信的投遞成功率，研究結果發現Random Forest為所有模型中表現最佳者，而掛號的屬性資料及投遞當下的雨量確實為影響投遞成功率的重要因子

資料分析流程圖


開始
研讀郵務流程以增加背景知識
將競賽資料放入資料庫
結合關聯表格(郵件狀態代碼表、未投妥原因代碼表等)
題目發想-找出影響投遞成功率的主要因素
資料擷取-擷取特種郵件追蹤查詢資料(TT表)中部份樣本
EDA-研究TT表中與投遞失敗相關之欄位
建模準備-選取TT表中狀態完整之掛號郵件
建模準備-與ACC表格中的欄位結合
建模準備-結合都市名、雨量等外部資料
建模準備-資料前處理及編碼
模型選擇及建立模型
模型預測結果分析及解讀
發想模型優化方案
研究結果資料視覺化
結束

EDA分析結果(一)

高達12.06%的掛號號碼曾投遞失敗

表一、母體資料中，掛號在投遞的過程中最常出現的代碼一覽

代碼	解釋	出現之不同掛號號碼筆數	佔所有掛號筆數的比例
Z4	運輸途中	79,316,299	89.06%
Y4	郵件投遞中	78,558,222	88.21%
I4	投遞成功	72,868,506	81.82%
A1	交寄郵件	67,214,789	75.47%
H4	投遞不成功	10,737,586	12.06%
G2	到達支局招領中	6,225,197	6.99%
Z2	招領郵件轉運中	3,542,594	3.98%
I2	投遞成功	2,775,145	3.12%
Y1	寄存送達郵局轉運中	2,292,774	2.57%
A2	交寄郵件	1,943,823	2.18%
Z1	信箱／代辦所郵件轉運中	1,918,001	2.15%
G1	到達支局信箱	1,774,986	1.99%
…	…	…	…
IL	文書退回投遞成功	6	0.00%
YL	(無解釋)	2	0.00%
HL	文書退回投遞不成功	1	0.00%

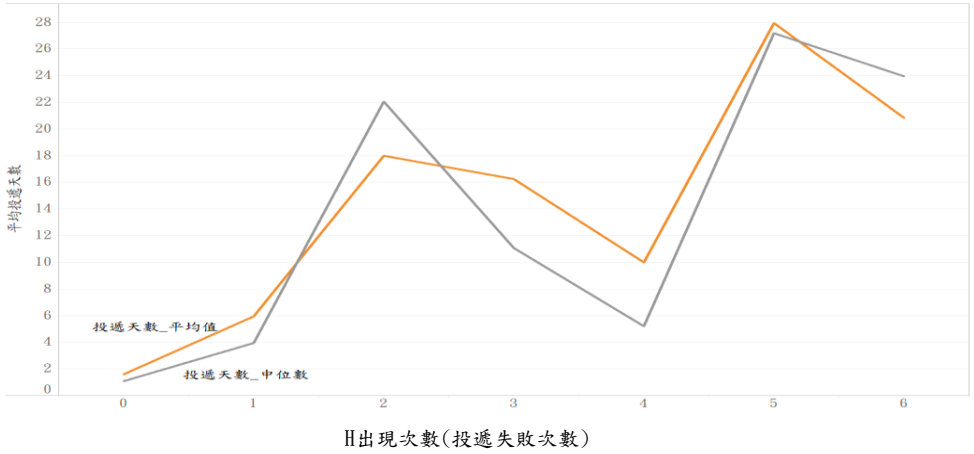
首先觀察母體資料的所有掛號號碼在投遞的過程中，最常出現的代碼(表一)，前幾名的代碼：Z4-運輸途中(89%)、Y4-郵件投遞中(88.21%)皆屬投遞過程中正常會出現的代碼

然而，在母體8900萬筆不同的掛號號碼資料中，有1074萬(12.06)%筆不同的掛號曾經投遞失敗，比例相當高

EDA分析結果(二)

投遞失敗的掛號郵件，會產生高達數十天之額外時間處理成本

表二、投遞失敗次數和總投遞時間分布圖



H出現次數(投遞失敗次數)	0	1	2	3	4	5	6
掛號號碼筆數	84,196	3,632	7,647	393	594	11	336
投遞天數_平均值	1.6	5.9	18.0	16.2	10.0	27.9	20.8
投遞天數_中位數	1.1	3.9	22.1	11.1	5.2	27.2	24.0

在有12.06%的掛號曾經投遞失敗此一高失敗率的情形下，為探討投遞失敗造郵務流程之時間影響，我們**隨機抽樣**約10萬筆掛號做進一步分析

由表二可發現，在投遞順利無投遞失敗(投遞過程中不曾出現H開頭的代碼)的情形下，總投遞時間的平均天數為1.6天，而在有投遞失敗的情形下(統計同一封掛號中出現H開頭的代碼的次數)，總投遞時間的平均天數提高至5.9天至數十天不等

投遞失敗對郵務流程影響甚大，不但讓投遞時間加長、更需額外人力及流程處理，使整體成本提升。再度呼應我們找出影響掛號郵件投遞成功率的因素，以發想方案提升投遞成功率，是重要且必須解決的議題

EDA分析結果(三)

投遞失敗的原因以「不在」及「按鈴無回應」為主

表三、有投遞失敗的郵件之未投妥原因列表

未妥投原因說明	筆數	佔投遞失敗郵件%
不在	13,319	53.6%
按鈴（呼叫）無回應	4,639	18.7%
查無此人	2,120	8.5%
無此地址	1,539	6.2%
遷移	1,155	4.6%
地址欠詳（錯誤）	580	2.3%
拒收	295	1.2%
查無此公司	239	1.0%
…	…	…
i郵箱已滿	1	0.0%

為了找出投遞失敗的原因，我們將樣本資料中曾有投遞失敗紀錄(狀態代碼中為H4)的掛號郵件獨立出來，觀察該批郵件的「其他欄位」中的註記，並結合「未投妥原因代碼表」，以找出掛號郵件投遞失敗的主要原因，將結果呈現於表三

我們觀察到未投妥原因以收件人「不在」(佔53.6%)及郵差「按鈴(呼叫)無回應」(佔18.7%)為主，因收件者不在家而使郵差投遞撲空為造成投遞失敗的主要因素

EDA分析結果(四)

投遞失敗的機率以中午11-14點最高；星期對投遞失敗率無影響

表四、投遞成功/不成功比例按投遞小時分

投遞時段(小時)	投遞狀態	所佔郵件數比例(%)	總投遞數
9	投遞成功	84.86	1,896
	投遞不成功	15.14	
10	投遞成功	85.3	2,456
	投遞不成功	14.7	
11	投遞成功	67.55	4,928
	投遞不成功	32.45	
12	投遞成功	57.54	14,164
	投遞不成功	42.46	
13	投遞成功	71.58	25,534
	投遞不成功	28.42	
14	投遞成功	78.77	22,499
	投遞不成功	21.23	
15	投遞成功	85.03	15,526
	投遞不成功	14.97	
16	投遞成功	88.82	10,941
	投遞不成功	11.18	
17	投遞成功	94.77	7,758
	投遞不成功	5.23	
18	投遞成功	96.89	3,762
	投遞不成功	3.11	
其它投遞時段	投遞成功	91.23	3,746
	投遞不成功	8.77	

表五、投遞成功/不成功比例按投遞星期分

星期	投遞狀態	比例(%)	總筆數
一	投遞成功	76.53	21,157
	投遞不成功	23.47	
二	投遞成功	78.67	23,672
	投遞不成功	21.33	
三	投遞成功	77.68	22,244
	投遞不成功	22.32	
四	投遞成功	77.87	22,426
	投遞不成功	22.13	
五	投遞成功	78.09	22,042
	投遞不成功	21.91	
六	投遞成功	95.87	1,647
	投遞不成功	4.13	
日	投遞成功	61.54	52
	投遞不成功	38.46	

為研究投遞時間是否影響投遞成功率，我們將樣本資料中的掛號號碼，將其在投遞成功(I4)/失敗(H4)的狀態發生的時間分別依照小時及星期分配，結果呈現在表四及表五

由表四可發現，投遞失敗率在中午11-14點明顯高於其他時段，其中在12點的投遞失敗率更高達42%。重覆隨機抽樣幾次(確保非特定樣本之偏差)皆得到類似結果。中午時段收件人外出覓食可能是造成投遞失敗的原因之一

由表五可發現，在週一至週五投遞失敗率皆在21%-23%之間，無顯著差異。週末因僅特定郵局投遞，不屬於正常投遞工作日，故不納入分析範圍。投遞星期對投遞成功率無明顯影響

處理環境

- 硬體:
  - 中央處理器: Intel Xeon CPU, E7530 1.87GHz 1.86GHz
  - 記憶體: 64.0GB
  - 作業系統: Windows 10 企業版
- 資料庫: PostgreSQL Database
- 資料擷取: Standard SQL
- 資料清理: Python 3.7
- 軟體-資料視覺化: Tableau 2018.1
- 所用之資料集
  - 特種郵件追蹤查詢資料(TT表)
  - 郵件狀態代碼表
  - 處理方式代碼表
  - 未投妥原因代碼表
  - 收寄資料明細檔(ACC表)
  - 外部資料：中央氣象局各都市氣象站逐日雨量資料

Source: <交通部中央氣象局:每日雨量，資料期間:2018/01/01– 2018/03/31，https://www.cwb.gov.tw/V8/C/D/DailyPrecipitation.html>

模型目的 & 資料清理(I)

模型目的

透過結合 掛號郵件的屬性資料(郵件種類、速度別等) + 寄達地資訊(都市名、是否位於都會區) + 寄達地雨量資訊 → 預測 → 每一封掛號號碼最終是否投遞成功

資料清理流程

步驟一、選取TT表中狀態完整的資料

在抽樣觀察TT表的資料時，我們發現部分掛號郵件中存在「狀態不完整」的現象，例如某掛號第一個狀態非「交寄郵件」而直接是以「到達支局信箱」為開頭；或者結尾狀態為「轉運中」這種停留在異常狀態的結尾。因此我們定義了正常的開頭及結尾狀態，並於資料庫中選取其中不同掛號號碼資料作為第一步資料篩選

- 正常的開頭狀態代碼: A1, A2, A3(交寄郵件)，排除無交件紀錄的掛號號碼
- 正常的結尾狀態代碼: H7& HL(退回投遞不成功), I開頭之代碼(投遞成功)

步驟二、將結尾狀態依照是否投遞成功分類

將取出的狀態完整的掛號結尾代碼依照最終投遞給收件人是否成功分成2類，新增「 is_success」欄位

- 投遞給收件人成功(is_success=TRUE): I1 & I2 & I4(投遞成功), I3(i郵箱取件成功), I5(收受人領取), I6(警局寄存送達)
- 投遞給收件人不成功(is_success=FALSE): H7& HL(退回投遞不成功), I7, I8, I9, IL(退回投遞成功)

步驟三、將TT資料與ACC表中的屬性資料結合

在資料庫上將從TT內狀態完整之50萬筆不同掛號號碼與ACC表中掛號號碼起號=迄號的郵件欄位結合起來：因為若起號≠迄號，代表同時交寄多個掛號號碼，無法分別判斷大宗交寄的每個掛號號碼屬性資料是否相同，故排除之。結合完畢後，產出約10萬筆不同掛號號碼的屬性及投遞狀態總表，且曾經投遞失敗的筆數佔樣本總筆數12%，與母體佔比相近

資料清理(II) & 合併資料集

步驟四、結合都市及鄉鎮區資料

因為想觀察投遞成功率是否存在著縣市間的差異(例如特定縣市投遞成功率較低)，因此我們透過ACC28的寄達郵遞區號，找出寄達地所在都市及鄉鎮區

步驟五、結合雨量資料

最後，為觀察投遞當下的雨量對投遞成功率是否有影響，我們透過ACC28的寄達郵遞區號，找出寄達地所在都市後，我們利用中央氣象局各都市氣象站逐日雨量資料及TT表的處理日期欄位(TT3)，結合出投遞當下的雨量

步驟六、將預測結果分類並定義預測欄位

我們將想要預測的掛號郵件依投遞狀態分成4個族群(Group)

Group	定義	說明
1	第一次投遞就成功且最終投遞成功	第一次投遞時的狀態為I1-I6，且之後沒有出現投遞失敗等異常紀錄
2	第一次投遞成功但最後失敗	第一次投遞時的狀態為I1-I6，但之後出現投遞成功被註銷(V2)等異常狀態
3	第一次投遞失敗但最後投遞成功	第一次投遞時的狀態為H4，但最後結尾為I1-I6(is_success欄位=TRUE)
4	第一次投遞失敗最後也投遞失敗	第一次投遞時的狀態為H4，最後結尾為H7, HL, I7, I8,I9,IL, is_success=FALSE

最終資料欄位一覽

預測用之因子欄位(feature)　　想預測的結果

欄位名稱	Mail_ID	ACC01	ACC02	…	ACC38	postal_code	city	District	is_city	rainfall	hour	Group
欄位說明	掛號號碼	ACC中的相關欄位				郵遞區號	城市名	鄉鎮市區	五都	降雨量	投遞時段	投遞狀態
	MailID_01											1
	MailID_02											2
	…											3
	MailID_N											4

資料前處理及資料編碼

步驟一、欄位處理	刪除結合後的資料集當中意義重覆的欄位
步驟二、空值處理	• 類別型數據，依該因子特性和數量，補為最常出現的類別(眾數)、或自成一類 • 數值型數據，以該因子平均數填補
步驟三、數據轉換	對類別型數據進行編碼處理
步驟四、特徵縮放	針對部分模型，為避免被值域較廣的因子主宰，將因子進行標準化
步驟五、切訓練集&測試集	將70%數據放入模型做訓練，其餘30%做測試，以驗證模型的表現

模型選擇

因為我們挑選的特徵(feature)欄位中多為屬性(categorical) 資料，且最終要預測的結果為分類類型結果(classification)，因此選用以下5個適合處理分類問題的模型進行預測，並將模型名稱及選擇原因列在表六

表六、模型及選擇原因

模型名稱	選擇原因
Logistic Regression	傳統上適用於預測分類問題的模型
kNN	最近鄰居法，透過向量空間距離中類別距離最小、頻率最高的方法進行分類，對於異常值較不敏感
Naive Bayes	計算在每個特徵值下，預測結果發生的機率。適合處理多類別問題
Decision Tree	將每個特徵值表達成一個節點，較容易根據模型結果表達出邏輯推論
Random Forest	對多變量預測表現佳，並且對存在缺失及資料不平衡的數據集包容力較強，較不易發生overfitting

模型預測結果- Random Forest表現最佳

表七、各模型預測準確度

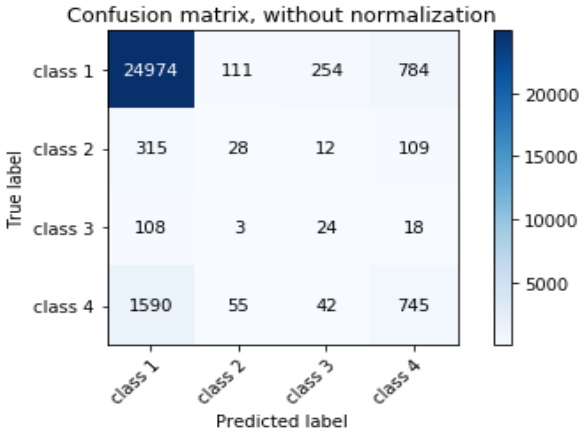
模型名稱	Accuracy	Precision
Random Forest	**90.7%**	**74.3%**
Decision Tree	90.6%	63.5%
kNN	90.1%	61.1%
Naive Bayes	88.0%	59.2%
Logistic Regression	86.2%	53.4%

我們以Accuracy及Precision作為衡量模型表現的指標

- Accuracy: 模型正確預測的比例
- Precision: 預測為某類別之正確預測的比例

由表七可發現，Random Forest在Accuracy跟Precision中預測準確度都最高，為所有模型中表現最好的，而Decision Tree次之

表八、Random Forest的混淆矩陣



由表八細看Random Forest模型的混淆矩陣(confusion matrix)來觀察4個族群中的預測值與實際值的差異，並綜合表七的結果來看，可發現Random Forest雖然在模型正確預測的比例及族群一(第一次就投遞成功且最後投遞成功)的預測度不錯，然而對於族群2, 3, 4之預測表現還可再加強

Random Forest中較重要的因子-屬性及降雨量對投遞成功率影響較大

表九、Random Forest模型中重要性較高之因子一覽

重要性	因子(feature)名稱	Feature Importance
1	ACC16 是否回執	0.354
2	ACC17 是否存證	0.139
3	ACC27 郵件種類名稱	0.096
4	ACC06 交寄日期	0.081
5	Rainfall 降雨量	0.073

我們以entropy衡量模型中各因子的重要性。由表九可發現，是否回執及是否存證為模型中重要性最高的兩個因子，有點出乎我們預期。原因推論為「回執」及「存證」為重要信函種類，郵差會特別留意是否送達，因此對「最終是否成功送達」此一預測結果影響較大

而郵件種類名稱(限時掛號信函、普通掛號小包等)、交寄日期及投遞當日的雨量以上因子對於掛號郵件對於最終是否成功送達也具影響力。寄達地的縣市別及是否位於都市區此2項地理位置資訊，非模型預測的因子中重要性較高者，無法判斷對掛號郵件最終是否成功送達是否有影響

模型未來優化方向

1. 針對重要/關鍵/具高解釋力因子(feature)，詢問郵局之專家以了解對郵務流程的影響層面及因果關係，藉以加強對重要因子的說明，並發想對應之解決方案
2. 使用Under-sampling 或 Over-sampling 等方式，優化目前因各族群樣本數不一而產生之數據不平衡的情況
3. 嘗試不同的預測組合：納入「準時」之條件，依郵件類型及地區預測是否在公告的時間內送達，以優化模型預測之細緻程度，並能針對郵政業務目標(例如包裹同一郵遞區號內次日投遞，跨區在第二至三日投遞)做實際投遞/預測投遞準時率分析，並找出影響準時之重要因子，協助郵局優化準時投遞的流程

結論

掛號郵件的高投遞失敗率(12%)不但使投遞時間提升，更需額外程序及人力處理，是個重要且必須解決的問題。
透過探索式資料分析及模型預測，我們找出影響掛號投遞成功率的因子主要包含：

- 投遞時間(小時)
- 掛號的屬性資料(回執、存證信函、郵件種類名稱)
- 交寄日期
- 投遞當日的降雨量

後續我們除了將持續優化模型、提升模型預測準確度外，也會將「投遞是否準時送達」納入模型預測範圍內。
並以提升投遞「成功率」與「準時送達率」為兩大目標，發想解決方案