郵件投遞成功率預測模型及其應用

| 台大資管B隊 | :孫君傳、劉正宇、簡辰安
| 指導老師 | :魏志平教授

摘要

- **目的**: 提高郵件投遞成功率以增加郵務士及郵局整體工作效率

- **方法**:
  - 統計: 找出影響投遞成功率的因素, 其中以時間、星期幾和天氣最為顯著。
  - 建立預測模型: 發現時間是影響最大的參數, 透過建製模型讓郵務士在投遞前了解手上郵件投遞的成功率
  - 建立最佳化路線與排程規劃: 將寄件過程描繪成VRPWT問題, 配合預測的最適時間點, 計算出最佳的寄件路線與排程。

大綱

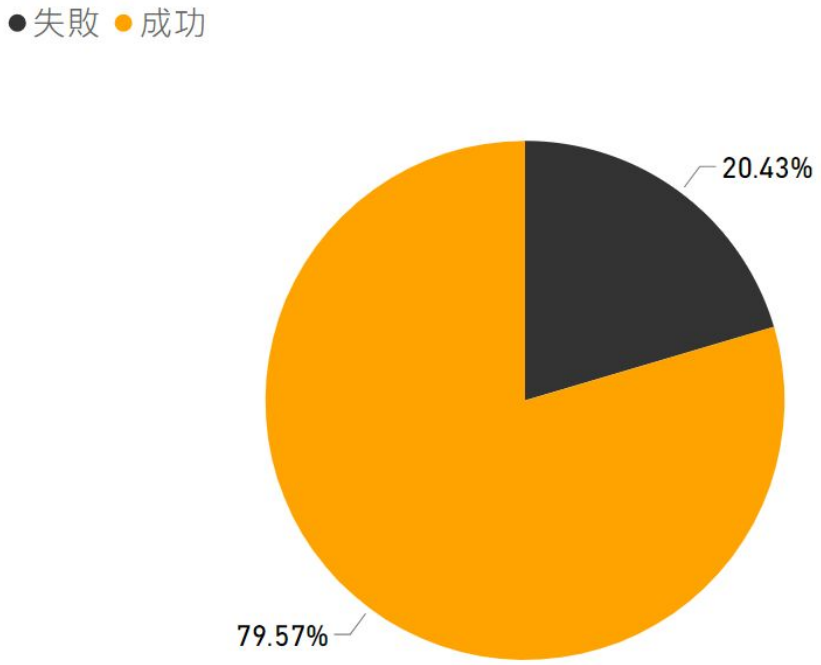
動機－回歸原點

中華郵政有著穩固的基礎客戶與規模，也一直致力於諸如i郵箱、民間通路、便利商店合作等創新，不過在新興業務仍面臨許多挑戰，對於企業客戶及電商配合上還有發展空間自是其中原因。

為了持續進步，我們認為中華郵政可以回到原點，**透過郵務數據的分析，提高郵件投遞的效率**，不只**縮短**了原先郵務士**送件及善後的時間**，更能解放這些人力資源，讓多出的人力可以投入其他業務，進而獲得發展新業務的**人力需求**，形成雙贏。

背景資料分析-郵件投遞成功率

- 由於現階段硬體設備的限制，我們選用高雄800584局號的資料作為這次報告的依據選擇此局號的原因為其處理的件數較多，且在這一季的運送路線較為規律。



- 總體1億以上郵件中，超過20%的郵件投遞失敗(如右圖)原因將在後面分為幾個部分討論:

1.投遞時段的影響

2.投遞日為星期幾的影響

3.投遞日天氣的影響

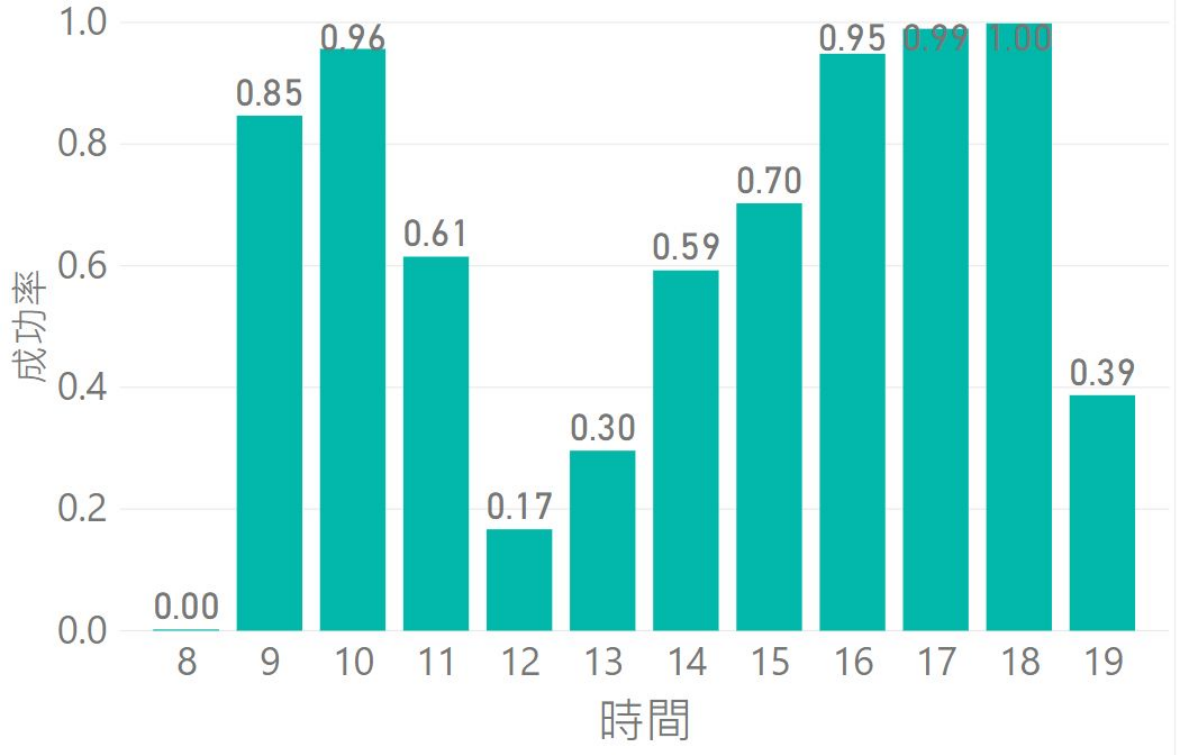
時段

在各個時段中的投遞成功率

右圖可發現時間是影響成功率的顯著原因

<u>高機率成功:</u>
<u>9點~11點、16~18點</u>
<u>高機率失敗:</u>
<u>8點前、12~15點、19點</u>

可能原因:
1.特定時段收件者比較不容易在家或無法收件
2.刷條碼習慣的問題，可能郵務士習慣在特定時間段刷投遞失敗的狀態



投遞日性質(星期幾)

在一個星期中各日的投遞成功率

由右圖可以發現到

<u>假日的成功率高於平日</u>

可能原因:
一般民眾假日通常會在家中, 容易簽收郵件

<u>星期四與五的收件成功率稍低於其他日子</u>

天氣

天氣是影響投遞的重要因素之一，從資料中可以看出:

雨量大到一定程度，投遞成功率有顯著下降

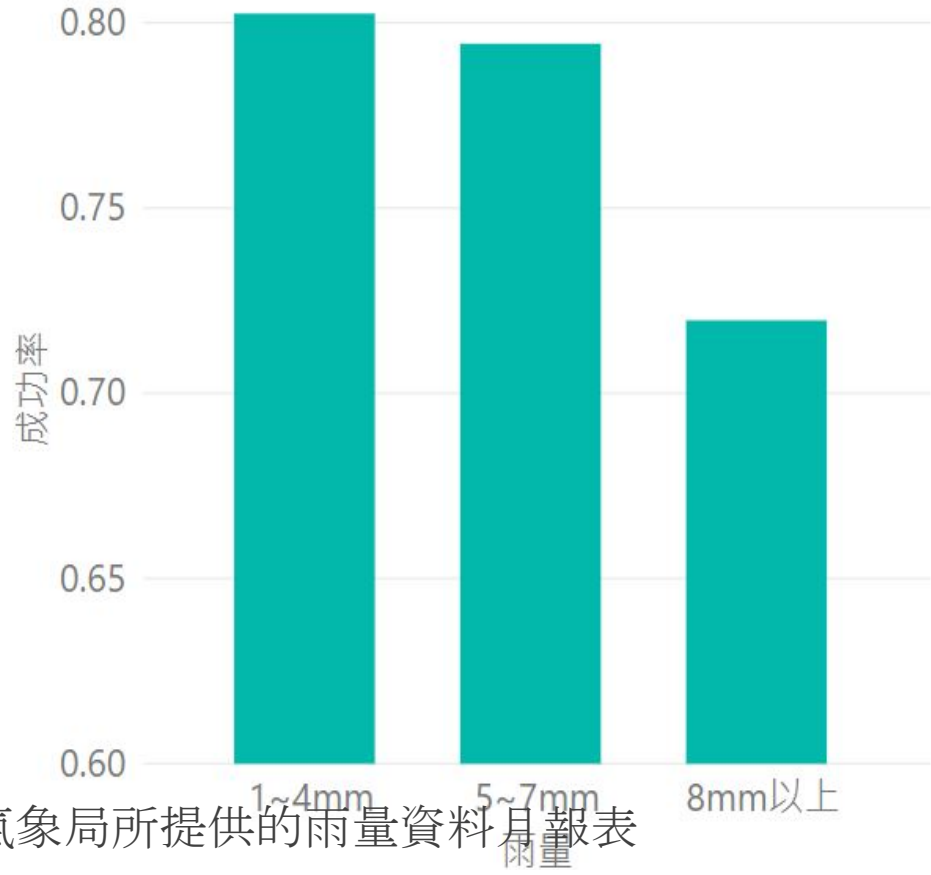
可能原因:

輕微下雨不太影響收件者與郵務士作息
當雨大到一定程度，便會影響到郵務士遞送的難易度與客戶的行為。

備註:

此資料集取自冬季的高雄，因此雨量較小
無法討論雨量很大的情況

在各種雨量中的投遞成功率



資料來源:中央氣象局所提供的雨量資料月報表

可能影響因素

我們希望找出統計證據，以證明**不同的因素確實會導致不同的寄件成功率**，如此便能蒐集相關變數，並建置相對應的預測模型。

Why 懷疑時間、日期等是影響寄件成功率的因素

How 使用 Pearson's Chi-Squared test of contingency table

What 找出統計證據，證明不同的變數會影響寄件成功率

Chi-Squared test of contingency table

不同日

H_0: 不同日的郵件寄送成功率　一樣

H_1: 不同日的郵件寄送成功率不一樣

```
            MON    TUE    WED    THU    FRI    SAT
Success  208824 234623 201855 226744 214015  16146
Fail      47348  45798  43499  74463  69320   2515

        Pearson's Chi-squared test

data:  (week)
X-squared = 11403, df = 5, p-value < 2.2e-16
```

```
        morning afternoon evening  night
Success   21297    236036  844864     10
Fail       5493    242072   32238   3140

        Pearson's Chi-squared test

data:  (time)
X-squared = 431940, df = 3, p-value < 2.2e-16
```

不同時段

H_0: 不同時間的郵件寄送成功率　一樣

H_1: 不同時間的郵件寄送成功率不一樣

Chi-Squared test of contingency table

不同雨量

H_0: 不同天氣(雨量)的郵件寄送成功率　一樣

H_1: 不同天氣(雨量)的郵件寄送成功率不一樣

by 中央氣象局所提供的雨量資料月報表

```
          1~4mm  5~7mm 8~10mm
Success 953147  56971  42157
Fail    234627  14750  16423

        Pearson's Chi-squared test

data:  (rain)
X-squared = 2385.9, df = 2, p-value < 2.2e-16
```

小結:

透過Pearson's Chi-squared test, 我們有充分的統計證據拒絕虛無假設, 也就是**不同時間點、日期和雨量所對應的寄件成功率並不相同**, 不論兩者間是相關還是因果關係, 這些變數都是我們建置預測模型的重要參考。

預測模型－投遞成功率

我們將**寄件成功與否**設為應變數，將更多可能影響的參數:**日期、時間、郵遞區號、氣溫、雨量、紫外線指數** 和 **同時間同區號的包裹數** 設為自變數，由於成功與否為**決定性問題(是/否)**，我們選擇使用 **logistic regression** 建置模型。

Why

了解參數對成功率影響的程度，提供規劃投遞路線的依據

How

使用R langauge 建置 Logistic Regression 模型

What

預測郵件在不同時間點、星期、天氣狀態下的投遞成功率

Logistic Regression

參數

	success	time	week	address	temp	rain	UVL	amount
2109356	I4	evening	WED	80700	19.6	1	9	234107
2109357	I4	afternoon	FRI	80700	15.0	1	7	234107
2109358	H4	afternoon	THU	80200	23.7	7	8	170119
2109359	H4	afternoon	WED	80200	24.0	1	10	170119
2109360	I4	evening	THU	80200	16.5	3	6	170119
2109361	I4	evening	WED	80700	21.1	1	11	234107
2109362	H4	afternoon	WED	80700	15.5	4	7	234107
2109363	I4	afternoon	FRI	80700	15.0	1	7	234107
2109364	H4	afternoon	THU	80700	23.5	1	7	234107
2109365	I4	afternoon	FRI	80700	15.0	1	7	234107

資料來源: 800584局號中郵件狀態代碼被登記為I4(成功)與H4(不成功)、且掛號號碼中有郵遞區號的投遞資料, 共約97萬筆, 每筆資料加上郵遞區號、當日平均氣溫、雨量、紫外線指數和同梯寄件數

時間:早上(9~11點)、下午(12~15)、傍晚(16~19)和晚上(剩餘時間)

訓練樣本: 7成的資料作訓練樣本, 剩下3成作比對

model

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9161   0.1958   0.2623   0.3314   3.5482

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.300e-01  1.789e-02 -18.450  < 2e-16 ***
timeevening  3.366e+00  6.497e-03 518.096  < 2e-16 ***
timemorning  1.520e+00  1.564e-02  97.174  < 2e-16 ***
timenight   -5.604e+00  3.149e-01 -17.793  < 2e-16 ***
weekMON      6.614e-02  8.498e-03   7.783 7.07e-15 ***
weekSAT      8.900e-01  2.656e-02  33.503  < 2e-16 ***
weekTHU      3.676e-02  7.763e-03   4.736 2.18e-06 ***
weekTUE      3.836e-01  8.262e-03  46.427  < 2e-16 ***
weekWED      6.146e-01  8.415e-03  73.040  < 2e-16 ***
temp        -1.748e-02  8.109e-04 -21.556  < 2e-16 ***
rain        -1.778e-02  1.047e-03 -16.970  < 2e-16 ***
UVL          2.409e-02  1.108e-03  21.731  < 2e-16 ***
amount       3.025e-06  2.892e-08 104.597  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1402502  on 1385149  degrees of freedom
Residual deviance:  942039  on 1385137  degrees of freedom
AIC: 942065

Number of Fisher Scoring iterations: 7
```

檢查模型

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9161   0.1958   0.2623   0.3314   3.5482

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.300e-01  1.789e-02 -18.450  < 2e-16 ***
timeevening  3.366e+00  6.497e-03 518.096  < 2e-16 ***
timemorning  1.520e+00  1.564e-02  97.174  < 2e-16 ***
timenight   -5.604e+00  3.149e-01 -17.793  < 2e-16 ***
weekMON      6.614e-02  8.498e-03   7.783 7.07e-15 ***
weekSAT      8.900e-01  2.656e-02  33.503  < 2e-16 ***
weekTHU      3.676e-02  7.763e-03   4.736 2.18e-06 ***
weekTUE      3.836e-01  8.262e-03  46.427  < 2e-16 ***
weekWED      6.146e-01  8.415e-03  73.040  < 2e-16 ***
temp        -1.748e-02  8.109e-04 -21.556  < 2e-16 ***
rain        -1.778e-02  1.047e-03 -16.970  < 2e-16 ***
UVL          2.409e-02  1.108e-03  21.731  < 2e-16 ***
amount       3.625e-06  2.892e-08 104.597  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1402502  on 1385149  degrees of freedom
Residual deviance:  942039  on 1385137  degrees of freedom
AIC: 942065

Number of Fisher Scoring iterations: 7
```

Deviance residuals is centered at 0 (good)

Wald's test for all coefficients: all statistically significant

時間對成功率的影響較顯著

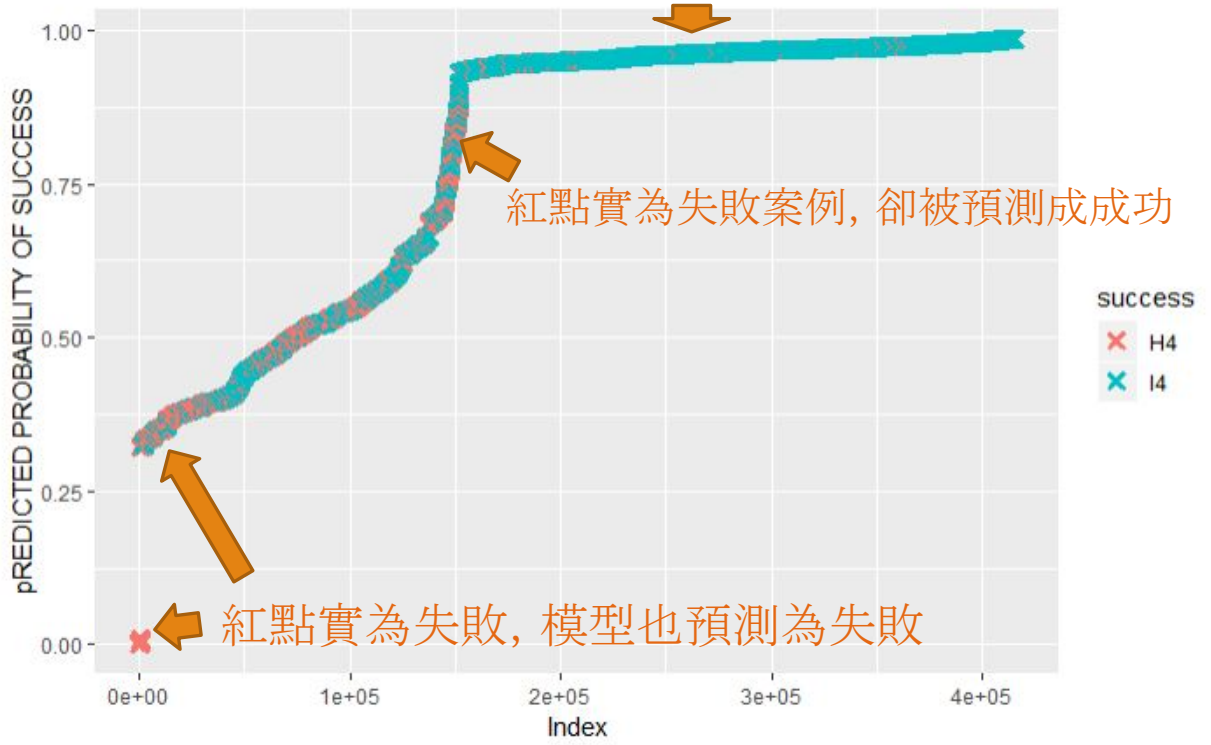
```
Confusion Matrix and Statistics

          Reference
Prediction     H4     I4
        H4  43693  30357
        I4  41189 300305

               Accuracy : 0.8278
                 95% CI : (0.8267, 0.829)
    No Information Rate : 0.7957
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.444

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5147
            Specificity : 0.9082
         Pos Pred Value : 0.5900
         Neg Pred Value : 0.8794
             Prevalence : 0.2043
         Detection Rate : 0.1051
   Detection Prevalence : 0.1782
      Balanced Accuracy : 0.7115

       'Positive' Class : H4
```

模型有82.8%的正確率

成功將成功案例預測成成功的準確度有9成

預測模型檢討

藍點實際為成功案例，模型預測為成功的機率值高達9成



此圖為測試資料被歸類為成功的機率值，同前頁模型的specificity所示，此模型對成功寄件的辨識度較高。

我們認為預測寄件成功率，**相當於預測客戶的行為模式**(寄件失敗多為送件時客戶不在)，由於一季的天氣變化不大、郵遞區號所內涵的地理範圍資訊太少，甚至郵務士的行動過於規律也會影響結果，這次**影響較大的變數只有寄件時間點和日期**(星期)。

要能提高預測準確度，除了嘗試不同模型如:SVM, random forest和MLP等，更重要的是讓訓練資料更具代表性。

我們建議若中華郵政希望**預測**不同時間點的寄件成功率時，應在**不違反客戶隱私權**的前提下，採用**長時間、具有客戶端的行為模式的代表性資料**，如更精確的**地理位置、去識別化用戶id、結合長期使用郵政服務**情形的統計和**推測人口結構**等資料，如此可以提高預測準確率。

預測模型結論

- 影響成功率的參數:**時間 > 星期 > 天氣 > 郵遞區號件數**(高雄冬天氣候過於穩定)
- 郵務士可以在投遞前了解個別郵件的投遞成功率。此預測模型判定為成功的郵件有高達**91%**可以投遞成功。對郵務士而言，可以更**專注在那些較能成功投遞的郵件**，避免延後投遞
- 可以透過模型了解**如何改變投遞模式**以增加投遞機率，舉例而言，可以在容易投遞成功的時間盡可能投遞郵件，並在容易失敗的時間進行其他業務，達到更好的時間分配
- 雖然現階段沒有，若**增加用戶id甚至地點資訊**，不只能加強模型的精確度，更可以進一步分析用戶或地點的特性，在正確的時間與狀態下將郵件投遞至特定位置，達到效率最大化
- 接下來會**結合最佳化模型**，進行更具規模的應用

預測模型應用:最佳化路徑模型

利用上述預測模型**預測郵件高機率能寄件成功的抵達時間**, 就能規劃一條有考慮對應投遞時間點的**最佳郵件投遞路線與排程**(最短、最快、最省錢依目標而異)

Why

找出如何**規劃投遞路徑以成功投遞最多郵件**

How

結合客戶郵址、道路資訊、過去的運送成本統計 和 **預測的最佳送貨時間**
便可以將寄件過程描繪成一個**整數線性規劃(ILP)的問題**

What

將寄送郵件的過程視為**Vehicle Routing Problem with Time Windows (VRPTW)**的問題
可以計算出成本花費最少、並分別在不同的最佳時間內送達至不同客戶的最佳路線。

VRPWT 模型

Obj:

$$Min \sum_{k}^{V}\sum_{i}^{N}\sum_{j}^{N} c_{ij}x_{ijk}$$

s.t. :

多物網路流基本式

$$\sum_{k}^{V}\sum_{j}^{N} x_{ijk} = 1, \quad \forall i \in C$$

$$\sum_{j}^{N} x_{0jk} = 1, \quad \forall k \in V$$

$$\sum_{i}^{N} x_{ihk} - \sum_{j}^{N} x_{hjk} = 0, \quad \forall h \in C, k \in V$$

$$\sum_{i}^{N} x_{i,n+1,k} = 1, \quad \forall k \in V$$

載重、車輛限制

$$\sum_{i}^{C}\sum_{j}^{N} d_i x_{ijk} \le q, \quad \forall k \in V$$

$$\sum_{k}^{V}\sum_{j}^{N} x_{0jk} \le |V|, \quad \forall k \in V, j \in N$$

寄件先後順序

$$s_{ik} + t_{ij} - s_{jk} \le M_{ij}(1 - x_{ijk}), \quad \forall i, j \in N, k \in V$$

最佳送件時間

$$a_i \le s_{ik} \le b_i, \quad \forall i \in N, k \in V$$

Sign constraint

$$x_{ijk} \in \{0,1\}, \quad \forall i, j \in N, k \in V$$

$$s_{ik} \in N^+, \quad \forall i \in N, k \in V$$

使用預測模型推測高機率寄件成功的時間區間

定義域

$V :=$ 郵差

$C :=$ 客戶編號

$N :=$ $|C| + 2$ 個節點，代表客戶端和分發站

$A :=$ 連接不同客戶的最短路徑

參數

$c_{ij} :=$ 從$\mathrm{i_{th}}$客戶到$\mathrm{j_{th}}$客戶的交通成本

$t_{ij} :=$ 從$\mathrm{i_{th}}$客戶到$\mathrm{j_{th}}$客戶的所需時間

$q :=$ 車輛載重/空間

$d_i :=$ $\mathrm{i_{th}}$客戶的計件重量/體積

$[a_i, b_i] :=$ $\mathrm{i_{th}}$客戶的最佳投遞時間

$M_{ij} := \max\{b_i + t_{ij} - a_j\}$(移動時間上限)

變數

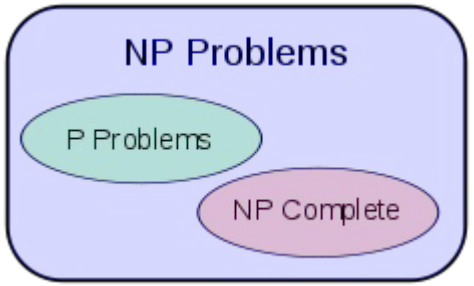
$$x_{ij}^k := \begin{cases} 1, & k_{th} \text{ 郵差是否從} \mathrm{i_{th}} \text{ 客戶端移動至} \mathrm{j_{th}} \text{ 客戶端} \\ 0, & o/w \end{cases}$$

$s_{ik} :=$ $\mathrm{k_{th}}$郵差開始服務 $\mathrm{i_{th}}$客戶的時間

最佳化路徑模型結論

- 目的:找出轄區內，不同郵差的最佳寄件路線(最快、成本最小或最多件因要求而定)
- 參數:客戶地點資訊、地圖資料、運送成本資訊、郵務士數量、最佳寄件時間
- 限制:
  - 目前資料缺乏關鍵性的**客戶地點資訊**
  - 此問題**(Integer Linear Programing)**本身為**NP-complete**問題，也就是說處理的數量越多，計算過程會越複雜，以至於使用低階配備會無法計算。
  - 對於**少量客戶**的郵遞區號(十萬戶)，可以透過solver (ex: **gurobi)直接運算**。
  - 如果**數量過多**，則必須透過Dynamic Programing、column generation等作法**事先簡化複雜度**，才能在合宜的時間內(天)計算。

GUROBI OPTIMIZATION | PREMIER PARTNER



結論

- 為了增進郵務士的投遞效率、減少郵務士及後續處理投遞失敗的時間，我們以位於高雄的局號800584的資料為例:
  - 驗證時間、星期、天氣.....等，找出具有影響投遞成功率的參數
  - 建立一個預測模型，能成功預測83%的投遞結果，對原本會投遞成功的郵件的預測更有91%，可根據模型在投遞前預測如何讓郵件投遞成功機率變高，進而改善投遞效率
  - 發展一個最佳路徑模型，可將預測模型的結果實際投入到應用層面(若是擁有地理及客戶資料，可提供更準確預估結果，最高將可以投遞失敗(20%)率降到最低)
- 未來發展:
  - 擁有更長跨度的資料:更精確的模型與參數分析
  - 使用更強的雲端運算:更複雜且完整的數據解析
  - 用戶id、地點參數:可建構完整的路線規劃模型，讓郵務士規劃投遞路徑時有強而有力的參考依據