

2023 臺北市捷運／公車轉乘 YouBike 每日租借量分析報告



組員

7113056014 施昶宇

7113056156 褚懿霆

7113056108 徐楚皓

7113056018 謝博勳

7113056162 廖彥瑋

# **主目錄**

[**主目錄** I](#_Toc199215850)

[**圖目錄** II](#_Toc199215851)

[**一、** **定義題目** 1](#_Toc199215852)

[**二、** **研究目標** 1](#_Toc199215853)

[**三、** **資料來源** 1](#_Toc199215854)

[**四、** **資料整理與觀察** 2](#_Toc199215855)

[**五、** **特徵值選取** 4](#_Toc199215856)

[**六、** **類神經網路的建模與驗證** 5](#_Toc199215857)

[**七、** **模型比較表** 7](#_Toc199215858)

[**八、** **結論** 7](#_Toc199215859)

[**九、** **程式碼連結** 7](#_Toc199215860)

# **圖目錄**

[圖 1 資料集內容 1](#_Toc199215865)

[圖 2 爬取出來的檔案內容 2](#_Toc199215866)

[圖 3 缺失值比例 2](#_Toc199215867)

[圖 4 出發量前十站分布圖 2](#_Toc199215868)

[圖 5 每日轉乘租借次數 3](#_Toc199215869)

[圖 6 星期別箱型圖 3](#_Toc199215870)

[圖 7 月份箱型圖 3](#_Toc199215871)

[圖 8 特徵值選取 4](#_Toc199215872)

# **定義題目**

預測任一天「全臺北市所有站點」因捷運／公車轉乘而產生的 YouBike 租借總量。

|  |  |
| --- | --- |
| 項目 | 說明 |
| 問題型別 | 單步時間序列迴歸 (Daily) |
| 目標變數 | daily\_rent\_count (整座城市每日轉乘租借量) |
| 商業意義 | 提前一天掌握需求 → 車輛再平衡、假日加班排班、行銷預算配置‧ 政策評估：大型活動 / 票價措施對轉乘量影響 |
| 成功判準 | 測試 MAE < 10 % 平均流量；R² > 0.8 |

# **研究目標**

* 預測任一天全市轉乘 YouBike 租借總量，以利營運調度、假日人力與行銷預算配置。
* 建立兩條基準模型——Dense 神經網路與 LightGBM——並比較效能。

# **資料來源**

資料集來自[台北市資料大平台](https://data.taipei/)，其中挑選[臺北市轉乘YouBike租借資料](https://data.taipei/dataset/detail?id=d51e8336-fceb-4275-a81a-2a246f4fca1b)作為本專題的研究資料，包含2023整個年度的資料。



圖 1 資料集內容

# **資料整理與觀察**

* 資料整理階段
  + 爬取12個月份的資料，並統整為一個CSV檔。



圖 2 爬取出來的檔案內容

* + 補缺失值(每個欄位都沒有缺值，所以減少這步)

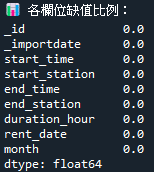


圖 3 缺失值比例

* 視覺化觀察階段
  + 出發量前十站，可以觀察到捷運公館站(2號出口)的出發人數最多，2023年約兩萬人。

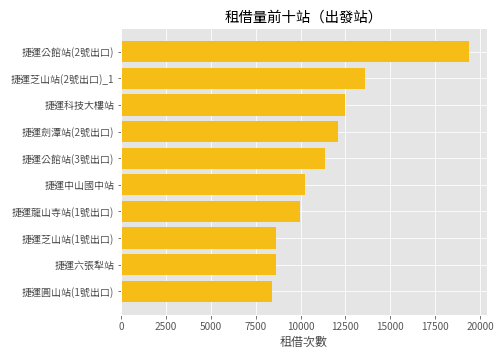


圖 4 出發量前十站分布圖

* + 每日總量折線圖，發現雨量較大（5 月梅雨、9 月颱風）出現局部低谷

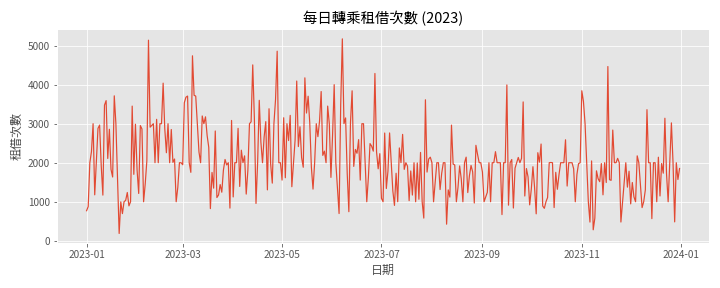


圖 5 每日轉乘租借次數

* + 星期別箱型圖，發現周末比較少人租借

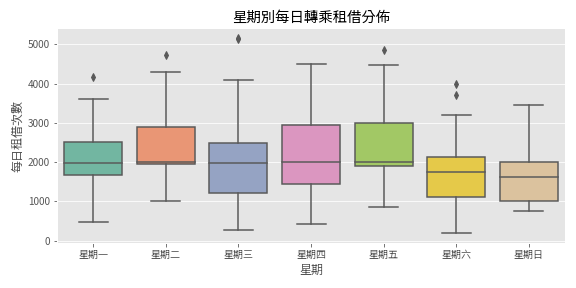


圖 6 星期別箱型圖

* + 月份箱型圖，九月波動最小

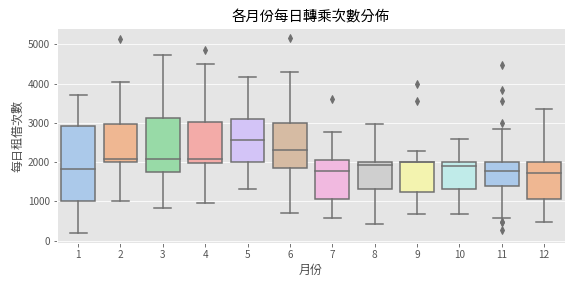


圖 7 月份箱型圖

# **特徵值選取**

* 1. 最小可用日曆集
  + is\_weekend ＋ day\_of\_year 兩個就能捕捉週期與年季節性，避免 One-Hot 過稀。
  1. 關鍵滯後
  + lag\_1/7/14 各對應「昨天」「上週」「雙週」，既解釋又容易讓模型抓到自相關。
  1. 多尺度趨勢
  + 7 / 14 / 30 天 rolling mean 讓模型知道「目前水平高低」與「短、中、⻑期波動」。
  1. 風險/不確定度
  + rolling std 提醒模型在高波動段降低過度擬合。

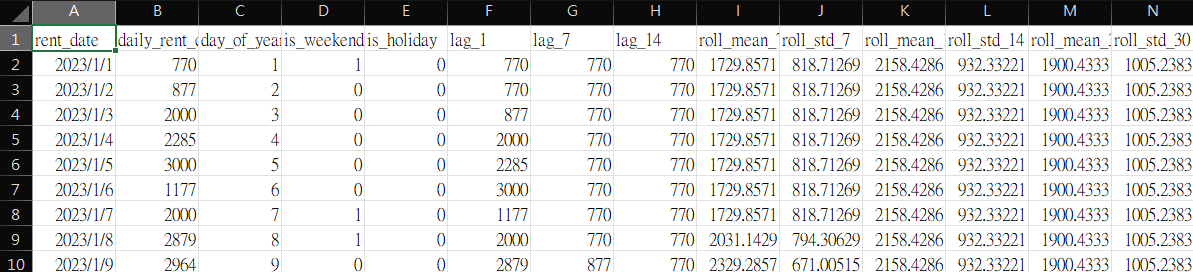


圖 8 特徵值選取

# **類神經網路的建模與驗證**

將資料分成前80%的時間(1-10月份)與後20%(11-12月份)的時間，其中前80%的資料會被輸入類神經網路建模，並利用後20%的時間來驗證你模型的有效性。

**此部分使用兩個不同的模型實作並做比較**

* + Dense Neural Network—架構細節

|  |  |  |
| --- | --- | --- |
| 區段 | 參數 | 說明 |
| **輸入層** | features = 16 | 13 個滯後/滾動特徵 + 2 日曆旗標 + 1 假日旗標 |
| **隱藏層 1** | Dense 64 units, ReLU | 高維投影，捕捉非線性組合；He 初始化 |
|  | Dropout 0.2 | 隨機棄權 20 % 神經元，抑制過擬合 |
| **隱藏層 2** | Dense 32 units, ReLU | 壓縮表示，提取關鍵特徵模式 |
| **輸出層** | Dense 1 unit (Linear) | 輸出單一連續值 ŷ |
| **優化器** | Adam, lr = 0.001 | 自適應學習率收斂快 |
| **損失函數** | MSE | 與 RMSE 指標一致 |
| **批次大小** | 16 | 在 300 筆序列×16 批次間取得平衡 |

**設計理念**：

* 用兩層 Fully‑connected 足以對「每日總量」這種低維度特徵做近似；
* ReLU 配 He 初始化可避免梯度消失；
* Dropout 與早停共同防止小資料過擬合。
  + LightGBM—架構細節

|  |  |  |
| --- | --- | --- |
| 模組 | 參數 | 作用 |
| **基學習器** | Gradient‑based One‑Side Sampling (GOSS)Leaf‑Wise Growth | 保留高梯度樣本以加速收斂；Leaf‑Wise 可在同樣深度下取得更低損失。 |
| **關鍵超參** | num\_leaves = 63 | 葉節點數 ≈ 模型複雜度上限；理論上 num\_leaves ≤ 2^{max\_depth}。 |
|  | feature\_fraction = 0.8 | 每棵樹隨機取 80 % 特徵，降低相關度、減少過擬合。 |
|  | bagging\_fraction = 0.8, bagging\_freq = 5 | 每 5 棵樹隨機抽樣 80 % 資料，再搭配 GOSS 取梯度前段樣本。 |
| **正則化** | 正則化 | L1 / L2 針對葉節點權重；對極長尾情況可加強。 |
| **學習率策略** | learning\_rate = 0.05 + 1 200 棵樹 | 小步長＋多樹可提升泛化；早停自動裁剪過多樹。 |
| **Early‑Stopping 回調** | lgb.early\_stopping(100) | 驗證集 100 棵無改善即停止；防過擬合。 |
| **評估回調** | lgb.log\_evaluation(200) | 每 200 棵列印一次 MAE；便於監控。 |

**演算法流程（簡述）**

1. **初始化**：以目標均值作為第一棵樹的預測 F₀。
2. **迭代 t = 1…T**：
   * 計算殘差梯度 gᵢ = ∂Loss/∂F\_{t-1}(xᵢ)。
   * GOSS 保留 |g| 最大的前 a% 樣本 + 隨機抽取 b% 的小梯度樣本 → 建樣本權重 wᵢ。
   * 在權重樣本上做 Leaf‑Wise 分裂，生成樹 h\_t。
   * 更新 F\_t(x) = F\_{t-1}(x) + η · h\_t(x)。
3. **early‑stopping**：若驗證 MAE 100 棵無改善即停止，留最佳迭代T\*.
4. **最終輸出**：ŷ = F\_{T\*}(x)。

# **模型比較表**

|  |  |  |
| --- | --- | --- |
| 指標 | Dense NN | LightGBM |
| MAE (筆/日) | 787.9 | **192.3** |
| RMSE (筆/日) | 1 085.4 | **307.2** |
| 1 − MAPE (%) | 68 % | **91 %** |
| ±10 % 容忍準確率 | 31 % | **78 %** |
| R² | 0.34 | **0.91** |

# **結論**

* 參數量並非越多越好：NN 雖參數少，但因序列樣本僅 365，仍易欠擬合；LightGBM 雖樹多，但 Leaf‑Wise + 早停可自動修剪冗餘。
* LightGBM 模型在無外部天氣的情況下已能將 MAE 壓低至 ~192 筆/日，對平均 2 500 筆/日的流量僅約 8 % 誤差，足夠做營運決策參考。
* 如果加入天氣因素之後，效果可能可以更好

# **程式碼連結**