Machine Learning Final Project Report

Team Member: 資管二 B11705051 陳奕廷 資管二 B11705058 廖振翔

Kaggle Team Name : 12345678

I. Introduction

- **Background Introduction:** 此次專案是要預測台北市多數 Youbike 的單車數量，以提升他們的單車調度系統。
- **Problem Statement:** 準確預測 112 個站點的單車數量。這個預測任務的焦點是在特定時間窗口內，每 20 分鐘的準確預估，而預測的內容則是與每個站點相關的非負整數值，即該站點的 Youbike 數量。

II. Data Collection and Preprocessing

- **Data Sources:** 資料來源是從教授發布的 final report 說明中的網址: https://github.com/hyusterr/html.2023.final.data/tree/release

  來自 *YouBike2.0 Taipei City public bicycle real-time information*。除此之外我們還使用了交通部中央氣象署的台北地區的每日累積雨量數據以及降雨時數，資料來源為: https://www.cwa.gov.tw/V8/C/D/DailyPrecipitation.html

- **Feature Selection:** 除了使用本次專案給予的資料，為了更精確的訓練及預測，我們還自己上網蒐集了台北地區的每日累積雨量及降雨時數，並且加上了每個日期是星期幾。以下是關於我們使用的 Feature 的示意圖:

date	rain_hour
20231002	0.6
20231003	8.2
20231004	13.5
20231005	7.6
20231006	0.4
20231007	2.1
20231008	2
20231009	1.2
20231010	0

date	rain
20231002	0.3
20231003	4.5
20231004	21
20231005	26
20231006	1
20231007	9.5
20231008	12
20231009	0.5
20231010	0

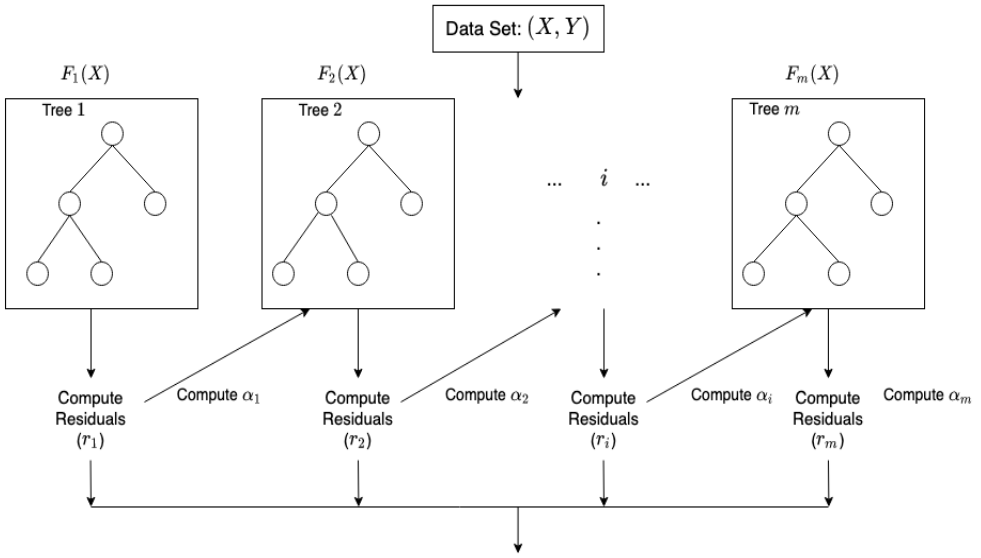
date	week
20231002	1
20231003	2
20231004	3
20231005	4
20231006	5
20231007	6
20231008	7
20231009	1
20231010	2

(由左至右分別為降雨時數、降雨量(mm)及星期幾)

III. Model Selection

- **Linear Regression**
  - Model Principles: Linear Regression 是一種基本的模型，透過建立目標變數（每 20 分鐘的單車數量）與一組預測變數（Features）之間的線性關係。模型的核心思想是通過擬合一條直線來進行預測，使預測值與實際值之間的差異最小化。
  - Implementation Details: 我們是使用 Python 中的 Scikit-learn 其中的 LinearRegression 模型進行訓練及預測。
  - Advantages: 線性回歸的優勢在於其簡單性和解釋性。易於理解和實現，適用於線性關係較為明顯的情況。計算效率高，不需要太多參數調整。
  - Limitations: 如果目標變數跟預測變數的關係是非線性的，那麼 Linear Regression 的預測會不準確，表現可能不如其他更複雜的模型。此外，它對異常值敏感，當數據中存在離群值時，模型的表現可能受到影響。
- **Random Forest**
  - Model Principles: Random Forest 是一種 Bagging 的學習方式，是透過建構多個基本的 Decision Trees，再以投票來提高整體模型的性能。每個 Decision Trees 都在隨機抽樣的子集上訓練，這有助於提高模型的泛化能力。
  - Implementation Details: 我們一樣是使用 Python 中的 Scikit-learn 其中的 RandomForestClassifier 模型進行訓練及預測。
  - Advantages: Random Forest 能夠處理大量特徵和高維數據，並且對於複雜的非線性關係有較好的適應性。
  - Limitations: Random Forest 模型相對複雜，訓練過程可能較為耗時，特別是在擁有大量 Decision Tree 的情況下。對於一些 Noise 較大的數據，Random Forest 可能會有 Overfitting 的問題。

- **XGBoost**
  - **Model Principles:** XGBoost（eXtreme Gradient Boosting）是一種 Gradient Boosting Algorithm，它在 Decision tree 的基礎上進行迭代，每次迭代都嘗試修正前一次迭代的模型的錯誤。這是通過計算梯度和標籤之間的殘差來實現的，然後將這些殘差作為新的目標進行建模。以下為漸變樹增強如何工作的簡要說明。



$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$$

where α_i, and r_i are the regularization parameters and residuals computed with the i^{th} tree respectfully, and h_i is a function that is trained to predict residuals, r_i using X for the i^{th} tree. To compute α_i we use the residuals computed, r_i and compute the following: $arg \min_{\alpha} = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}))$ where $L(Y, F(X))$ is a differentiable loss function.

  - **Implementation Details:** 我們是透過 python 中的 xgboost 來 import XGBClassifier 來進行訓練及預測。
  - **Advantages:** XGBoost 除了可以做分類也能進行迴歸連續性數值的預測，而且效果通常都不差。並透過 Boosting 技巧將許多較弱的 Decision Trees 集成在一起形成一個強的預測模型。
  - **Limitations :** 與其他複雜模型一樣，XGBoost 的解釋性相對較差，可能難以深入理解模型的內部工作。另外，它的訓練時間相對長，需要較多的計算資源。

IV. Implementation Details

- **Reading Data:** 我們把提供的資料都讀取進一個 dictionary 裡面，為了避免查詢時的麻煩，取資料時只需要提供車站名稱即可，這樣取資料會更準確。讀進去的資料如下圖。

```
entry = {
    'date': str(root)[-8:],
    'bike_stop': str(file)[:9],
    'time': time,
    'total': value["tot"],
    'current': value["sbi"],
    'can_park': value["bemp"],
    'open': value["act"],
    'rain_hour': rain_data[f'{str(root)[-8:]}'],
    'holiday': holiday_data[f'{str(root)[-8:]}']
}
```

  不過真正拿來 train 的資料不會使用到 date、open 還有 bike_stop，bike_stop 拿來當 dictionary 的 key 做尋找用途、open 則是決定要不要取用的標準，以免沒開放的車站影響 model 的訓練。此外，因為輸入的 time 是 24 小時制，我們把它轉換成分鐘以利訓練。

- **Reading Feature:** 我們是透過讀取 Excel 檔案來輸入自己蒐集的 Feature 資料，例如下雨時數跟星期幾，我們希望透過更多的 feature 來做到更精確的預測，同時我們也有使用過更精確的下雨時數或毫米數來做訓練，發現 validation error 很高，推測是模型 overfit，於是最後採用台北市下雨時數當作 feature 作使用。

- **Training Model:** 我們最後採用的是 XGBOOST 來做訓練，當中我們也使用幾個 validation error 表現不錯的參數來做 hyper parameter 的挑選，其中包含 n estimator 還有 learning rate。

```
Most common n estimator: 200
Most common learning rate: 0.1
most common pair: (200, 0.1)
```

  我們最後選擇了 200 和 0.1 當作我們後續做訓練時的參數。

- **Predicting:** 做 predicting 的時候，我們不僅需要使用專案提供的資料，還要去天氣網站調閱後一個禮拜的下雨時數報告，但也不能說非常準確，它們估計的有時跟現實差距也很大。

V. Performance Comparison

- **Evaluation Metrics:** 我們將使用 Mean Squared Error(MSE)、R-squared 來評估三種模型的效能。
- **Training Time:** Random Forest 跟 XGBoost 都需要耗費我們 4-5 個小時去訓練，但 Linear Regression 就跟老師所謂的一步登天一樣，很快就可以訓練完成，因此我也可以低成本的初步測試程式碼的正確性。
- **Prediction Time:** 跟訓練比起來就簡單很多，三種都需要耗費 20 分鐘左右。

VI. Details Of Trainging

- **Method:** 這次目標是要在某些車站中做腳踏車剩餘量的預測，考量到一車站跟車站間有蠻多的不同的，畢竟這涉及到地理位置等等的因素，像是離學校很近，或者是捷運站的交踏車站，那可能就會在下班、下課時有滿站或空站的情況發生，所以我們在每個車站上都訓練一個 model 出來，因此大約有一千三百多個模型，然後在預測的時候根據需要的車站來做相對應的預測。

```
Bike stop ID:  500108041
0.10324084012010477
Bike stop ID:  500108043
0.09950858221039416
Bike stop ID:  500108044
0.0926356900225966
Bike stop ID:  500108045
```

利用上述方法我就可以在每個車站都跑一個 validation error，當作我觀察 error 的標準。

VII. Model Selection Recommendation

在選擇最適合的模型時，我們最後採用的是 XGBoost 作為我們解決此次轉案的主要演算法。以下是我們推薦 XGBoost 此種演算法的理由:

- Outstanding Predictive Performance: 在這三個演算法中，我們透過 XGBoost 所獲得的模型是表現最好的，不論是在我們自己的 Validation 或是上繳 Kaggle 所獲得的成績中，XGBoost 都讓我們的錯誤率下降了很多，這是我們之所以選擇此演算法的主要理由。
- Handling of Noise and Missing Value: 我們的真實世界數據集中可能存在 Noise 或 Missing Value。也包括這次專案預測 Youbike 數量的情形，我們所採用的數據可能是有 Noise 跟 Missing Value 的。XGBoost 具有強大的 Noise/Missing Value 處理能力，能夠在模型訓練過程中有效地處理這些情況。
- Prevent Overfitting: XGBoost 內置了 Early Stopping 機制，這有助於防止模型在訓練過程中對訓練數據 Overfitting。這意味著我們的模型更有可能在新數據上表現出色。

VIII. Possible Improvement Directions:

- **Expanding Feature Set:** 可以多增加一些 Feature 的數量來提高預測的準確性，例如國定假日或大型事件可能會有更多的人群，交通狀況也可能會影響使用 Youbike 的意願，或是人口密度等等。
- **Utilize More Efficient Model:** 因為基本上整段程式碼最費時的部分是 Training 的部分，所以希望之後可以使用更有效率且更準確的 Model 進行訓練。

IX. Appendix

- Project Code: https://github.com/Tim4x4x4/ML
- Dataset Information:
  https://github.com/hyusterr/html.2023.final.data/tree/release
  https://www.cwa.gov.tw/V8/C/D/DailyPrecipitation.html
- Reference: https://arxiv.org/pdf/1603.02754.pdf
  https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/xgboost-HowItWorks.html
  https://ithelp.ithome.com.tw/articles/10247936