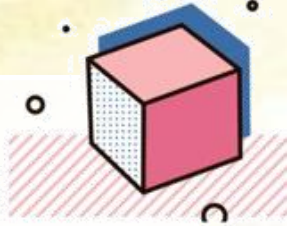


機器學習期中提案報告

筆電價格預測

班　　級：電通四甲

組　　別：第5組

報 告 者：許至佑、林垣志、許楷俊

指導教授：曾士桓

日　　期：112.04.24

Project Statement

- 電腦配置需求日益漸增
  - 擁有一台機體設備良好的電腦非常重要
- 市面上的硬體規格五花八門
  - 消費者對於價格與規格不了解，選 CP 值高就對了?
  - 欲找到最理想且合理價錢的款式便需要花很多時間做功課

Motivation and Challenge

- Motivation
  - 欲找到最理想且合理價錢的筆電
    - 市場利用平均價格計算硬體規格的價錢
    - 了解市場趨勢，進行價格比較，找到最好的交易
  - 利用機器學習的技術推估筆電價格
    - 運用 scikit-learn 套件中的機器學習模型
      - Linear Regression、Random Forest、XGBoost

Motivation and Challenge

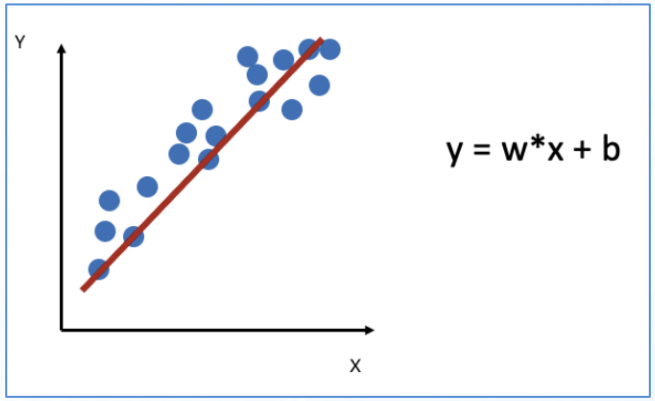
- Challenge
  - 資料前處理與特徵選擇
    - 缺失值與異常值的取捨
    - 選擇資料中有用的特徵
  - 使用 scikit-learn 套件的模型
    - 評估模型的選擇
    - 超參數設定

Related Work

- **資料集**
  - Kaggle網站 ➔ 「Laptop Price Prediction using specifications」
    https://www.kaggle.com/datasets/arnabchaki/laptop-price-prediction
- **方法**
  - Linear Regression
    [Python 實作] 迴歸模型 Regression – PyInvest (pyecontech.com)
  - XGBoost 論文
    https://arxiv.org/pdf/1603.02754.pdf
  - XGBoost 實作
    https://ithelp.ithome.com.tw/articles/10301273
  - Random Forest【分類、回歸】
    https://ithelp.ithome.com.tw/articles/10303882?sc=iThelpR

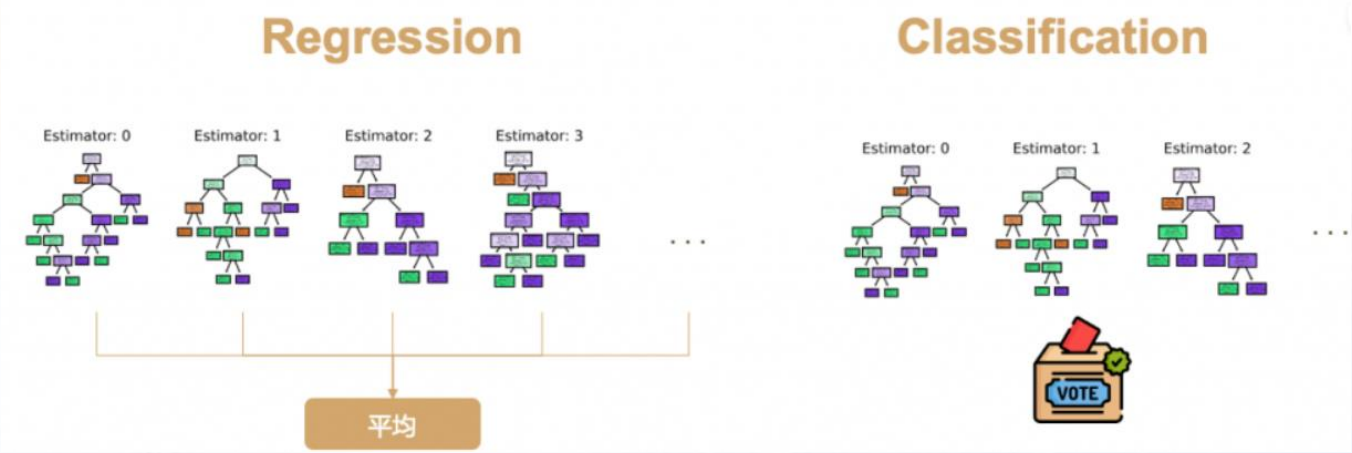
Related Work

- Linear Regression (線性回歸)
  - 是統計上在找多個自變數和依變數之間的關係所建出來的模型
  - 透過計算出迴歸模型線的方程式，就能獲得預測值
  - 若遇到 underfitting，可提高迭代次數解決或提高模型複雜度
  - 若遇到 overfitting，確認是否訓練資料太少或迭代次數過多
  - 能適時使用正規化 (L1/L2) 處理



Related Work

- Random Forest (隨機森林)
  - 由多棵決策樹所組成，進階版的決策樹
  - 使用 Bagging 加上隨機特徵採樣 (Bootstrap) 的方法所產生
  - 比較不易受極端值影響
  - 不足在於無法控制模型內部的執行

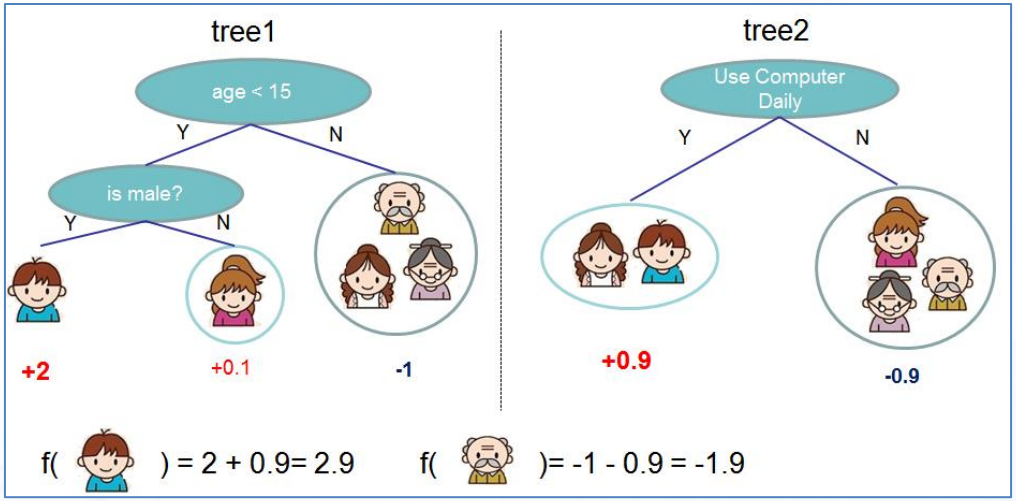


圖片來源：https://ithelp.ithome.com.tw/articles/10272586?sc=pt

Related Work

- eXtreme Gradient Boosting (Xgboost)
  - 基於 Gradient Boosting 做延伸
  - 結合 Bagging 和 Boosting 的優點
  - XGBoost 可想成是多個 CART 進行融合

dmlc XGBoost

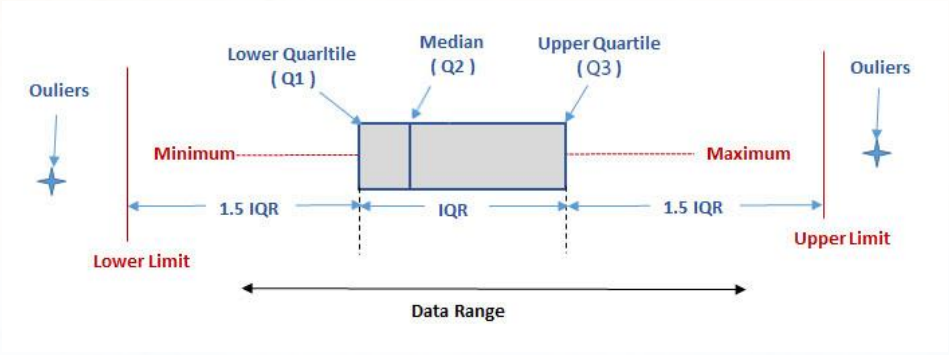


圖片來源：https://ithelp.ithome.com.tw/articles/10301273

Potential Solutions

資料前處理

- 離群值
  - IQR (四分位距)➔Q3 – Q1，移除 Outliers



- 缺失值
  - 直接刪除資料或單一特徵
  - Imputer 方法，像是 KNNImputer、Marking imputed values

Potential Solutions

資料前處理

- 資料型態轉換
  - One-Hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

  - Target Encoding

Fruit	...	Price
apple	...	50
banana	...	35
guava	...	95
banana	...	15
apple	...	30
guava	...	125

Fruit	...	Price
20	...	55
12.5	...	35
55	...	95
12.5	...	15
20	...	30
55	...	125

Avg.Price / times

圖片來源：初學Python手記#3-資料前處理(Label encoding、 One hot encoding) | by Pat Huang | Medium

Potential Solutions

特徵選取

- 使用 Sklearn 內建模型對特徵重要性評估
  - ➢ SelectFromModel()
- Pearson 相關係數
  - 衡量兩變數間「線性」關聯性的高低程度

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

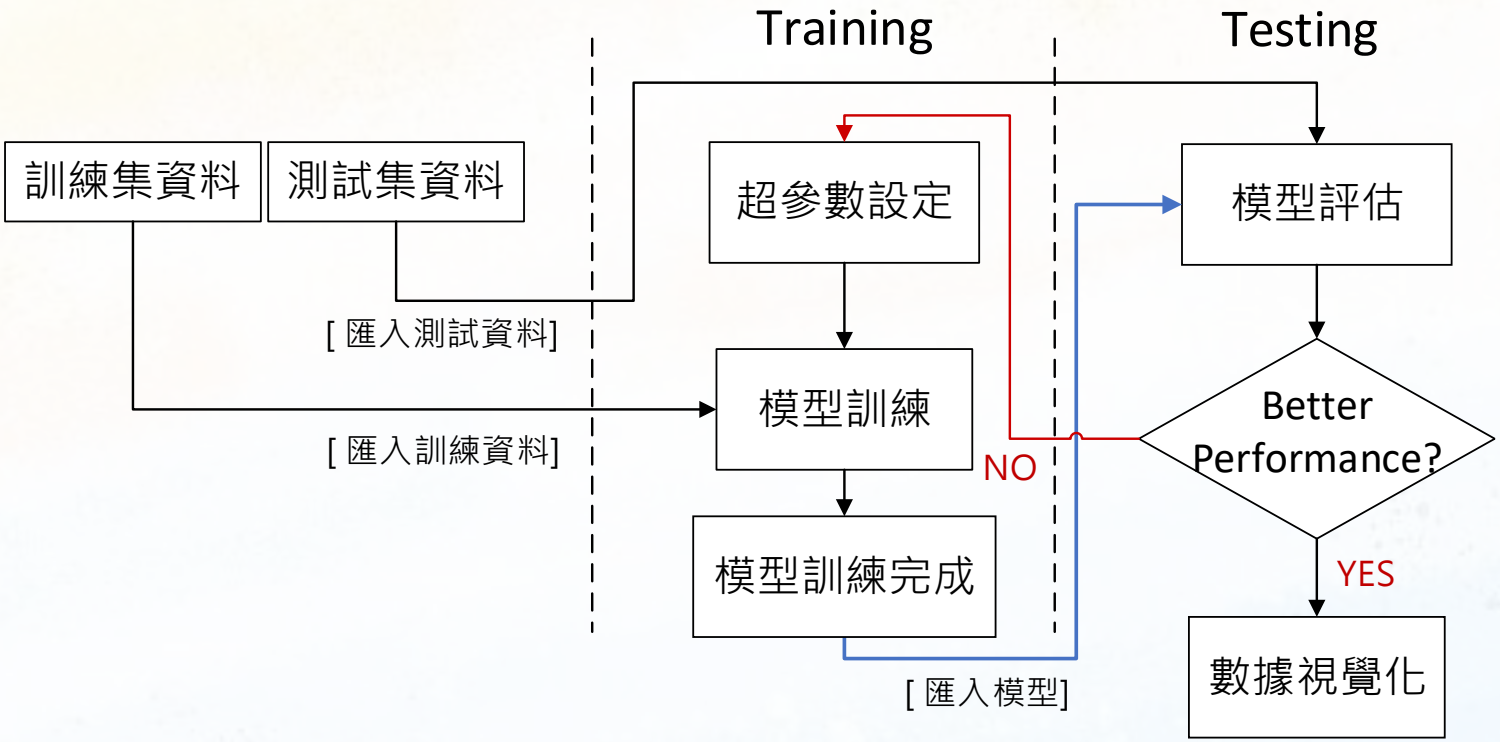
- CFS (Correlation-based Feature Selection) + BFS(Best First Search)

參考文獻：

1. Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.
2. Mohand, M. & Selamat, A. & Krejcar, O. & Fujita, H. & Wu, T. (2020). An analysis on new hybrid parameter selection model performance over big data set.

Potential Solutions

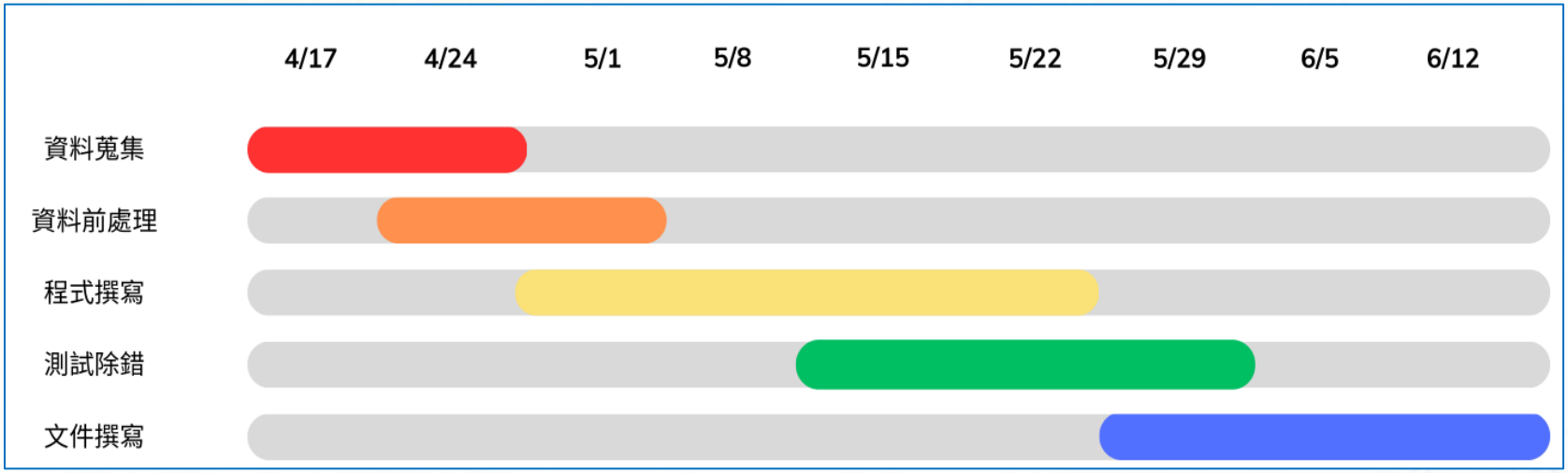
機器學習流程



Resource Required

- **Device and tools**
  - **硬體**
    - Windows 11 專業版 64位元
    - CPU：Intel(R) Core(TM) i5-13500 2.50GHZ
    - RAM：32G
  - **軟體**
    - 編譯工具：Spyder / Jupyter Notebook
    - python 3.7+、scikit-learn library、numpy、pandas、matplotlib
- **Human assignment**
  - 林垣志：資料蒐集、程式撰寫、報告
  - 許至佑：資料蒐集、程式撰寫、報告
  - 許楷俊：資料蒐集、程式撰寫、報告

Schedule



感謝聆聽

Plan B

- 主題：可回收塑膠瓶分類
- 使用方法：Pytorch、Numpy
- 資料集：
  Plastic Recycling Codes | Kaggle