AI.FINTECH
AI 金融科技中心

XGBoost 隔日股價漲跌預測-Python 實作

國立高雄科技大學金融資訊系教授兼AI金融科技中心主任
林萍珍

Shubham Malik, Rohan Harode, Akash Singh
XGBoost: A Deep Dive into Boosting
Updated Frebruary 2020
https://www.researchgate.net/publication/339499154_XGBoost_A_Deep_Dive_into_Boosting_Introduction_Documentation

致謝

- **部分圖片和內容來自以下人士和機構:**

Shubham Malik, Rohan Harode, Akash Singh
XGBoost: A Deep Dive into Boosting
Updated Frebruary 2020
Technical report

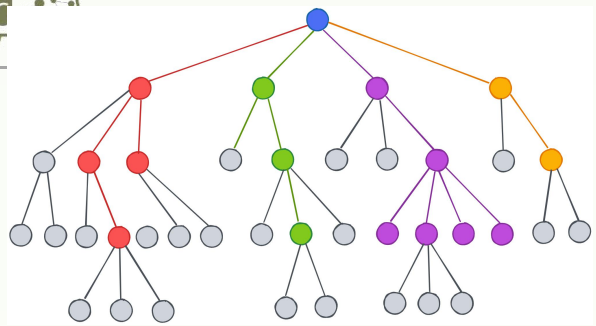
https://www.researchgate.net/publication/339499154_XGBoost_A_Deep_Dive_into_Boosting_Introduction_Documentation

https://zhuanlan.zhihu.com/p/584124751

XGBOOST 隔日股價漲跌預測

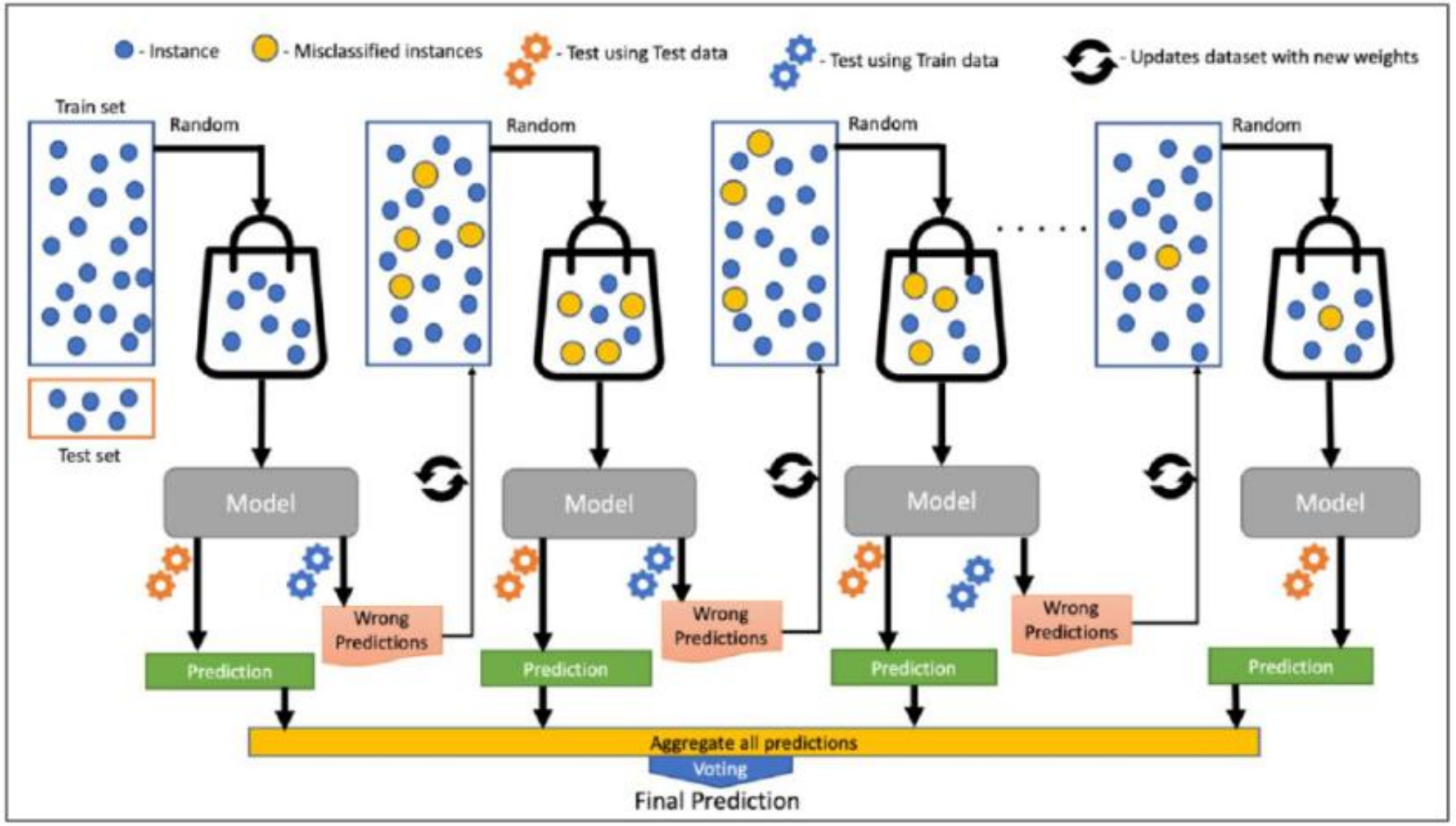
- **複合式學習(Assemble Learning)**
- 提升(Boosting)演算法運作原理
- **梯度提升Gradient Boosting圖解**
- XGBoost為何是機器學習首選算法?
- **決策樹的修剪 (Tree pruning)**
- SSR計算葉與樹的誤差平方合
- **程式碼講解**
  - **訓練、驗證、測試資料切割**
  - 參數設定:樹深度、學習率...
  - **創建XGBoost分類器**
  - 訓練XGBoost模型
  - **預測測試集**
  - 測試結果做混淆矩陣計算
- **AI 實作之參數與資料集的校調**

分類和回歸樹(CART)



- **Classification And Regression Trees (CART)**
  - 監督式機器學習算法，用於預測建模，多個獨立變數預測一個依賴變數（目標）
  - 類似樹的結構，頂部是根。
  - 分類(Classification): 當目標變量是固定類別(不連續)，這個算法被用來識別目標最有可能落入的類別，例如: 預測股市上漲或下跌兩類。
  - 回歸樹(Regression trees)：當目標變量是連續的，這棵樹/算法被用來預測連續的值，例如預測大盤指數, ex: 預測明天大盤是18000點。
- **XGBoost內部結構模型設計源自於CART，差別在於XGBoost是模型樹不是分類樹或回歸樹**
  - 模型樹的葉節點輸出值不是分到該葉節點的所有樣本點的均值（回歸樹），而是由一個函數產生的值。
- **XGBoost是機器學習一種非結構化梯度提升的演算法，改善CART準確度不足、損失高、結果變異大等問題。**
- **XGBoost可以處理回歸和分類問題，即可以預測實數(連續)也可以預測類別(不連續)。**

提升(Boosting)演算法運作原理



梯度提升Gradient Boosting

- **梯度提升是提升演算法(Boosting algorithm)的一種特例。**
- **依梯度下降(Gradient Descent)演算法最小化錯誤產生決策樹。**
- **梯度提升與梯度下降會根據錯誤更新模型（弱學習者）。**
  - **梯度提升藉由梯度下降的演算法來調整學習的權重。**
  - **此算法利用損失函數中的梯度-變化量的方向，迭代優化模型的誤差，藉此更新權重。**
  - **誤差指預測值和實際值之間的差異。**

$$w = w - \eta \nabla w$$
$$\nabla w = \frac{\partial L}{\partial w} \quad where\ L\ is\ loss$$

Gradient 做法是一階導數

w 代表向量的權重；η 是學習率

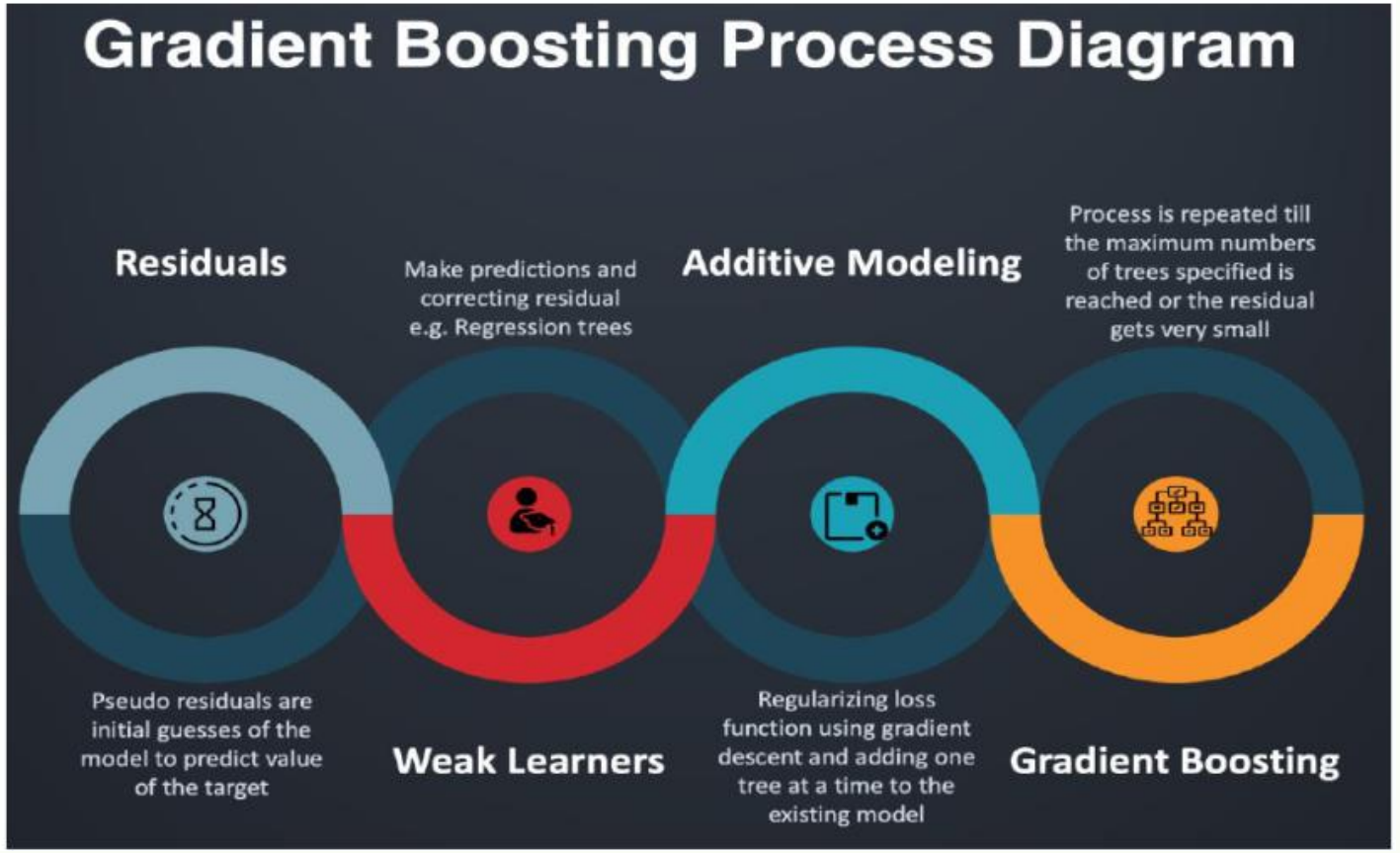
梯度提升過程圖解

殘差（Residuals）：對目標值進行初步猜測，結果為擬殘差。接著使用回歸樹進行預測並修正殘差。

弱學習者（Weak Learners）：透過加入弱學習者(一次加入一棵樹)，來對現有模型進行修正，並使用梯度下降來調整損失函數。

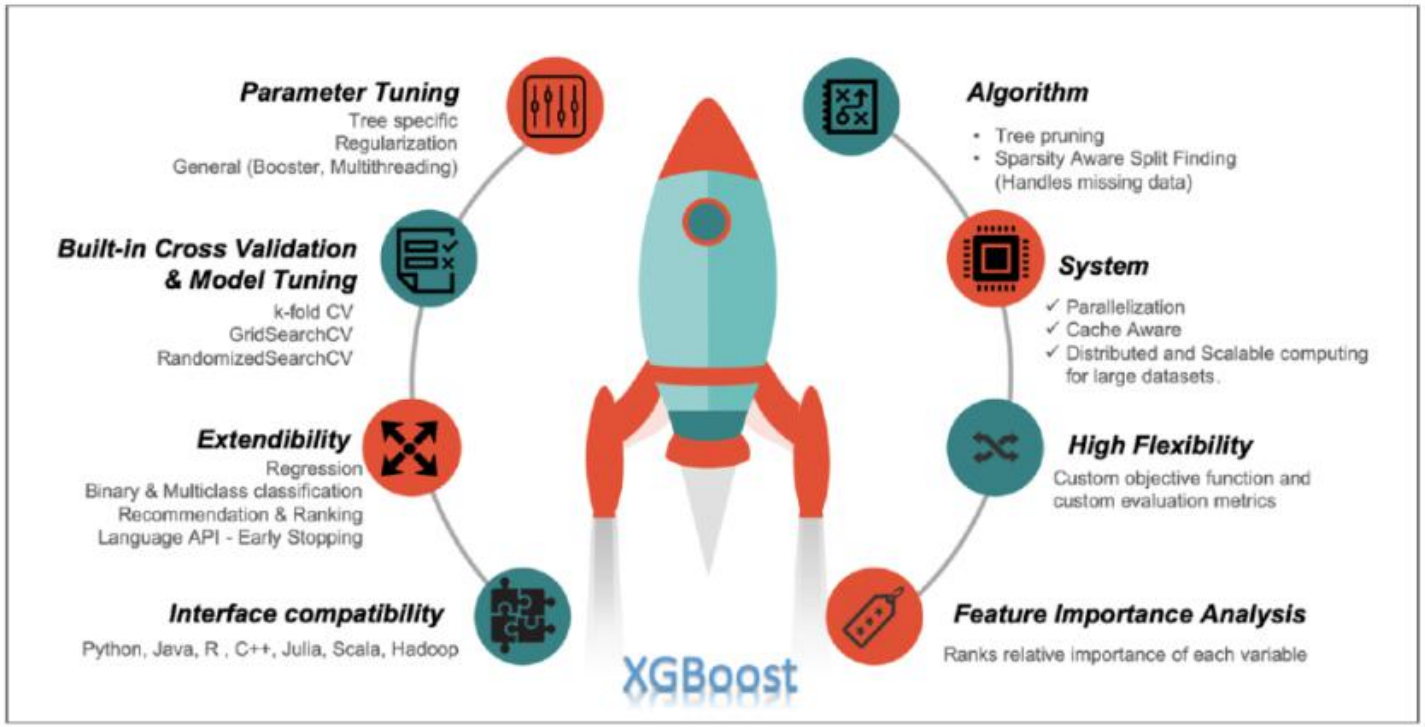
加法建模（Additive Modeling）：通過迭代添加弱學習者來不斷優化模型，每次添加為改善模型對數據的適應。

梯度提升（Gradient Boosting）：這個過程會持續重複，直到達到指定的樹的最大深度或殘差變得非常小為止。



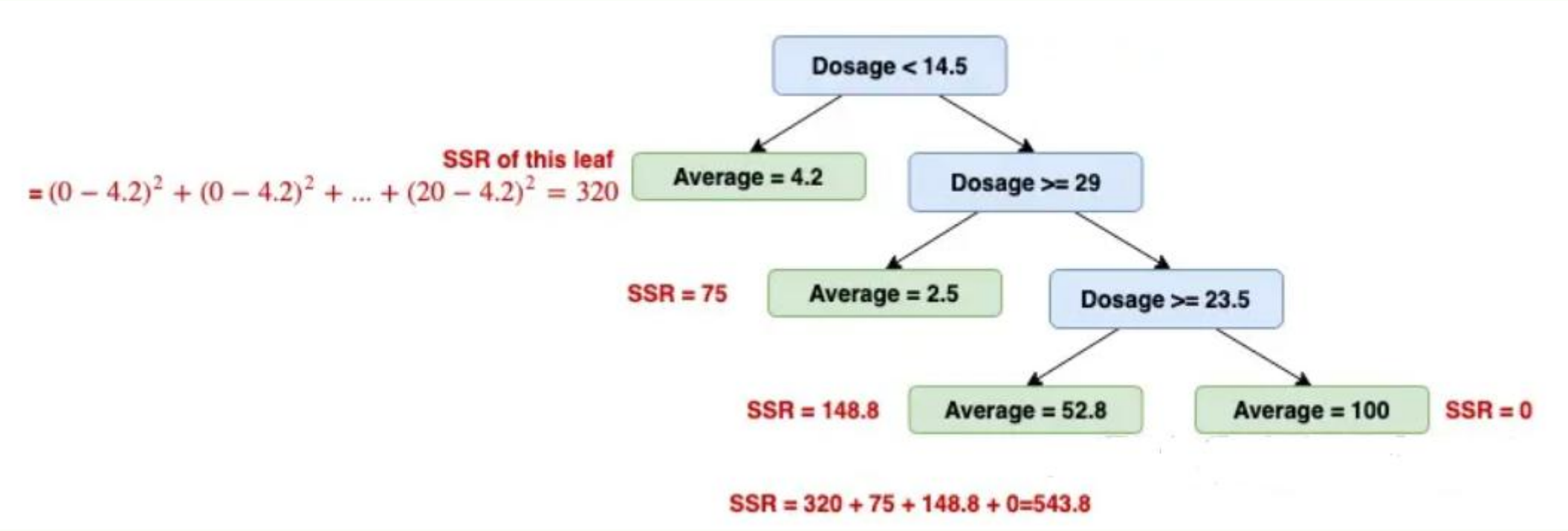
XGBoost為何是機器學習首選算法?

- **參數調整（Parameter Tuning）**:包含特定於樹的調整和一般性調整，比如提升方法和多執行緒。
- **內建交叉驗證與模型調整（Built-in Cross Validation & Model Tuning）**:支援k折交叉驗證、網格搜索和隨機搜索等方法來優化模型。
- **可擴展性（Extendibility）**:能夠處理二元和多類分類、回歸，並提供早停機制來防止過度配適。
- **介面兼容性（Interface compatibility）**:支持Python、Java、R、C++、Julia、Scala、Hadoop等多種程式語言。
- **系統（System）**:支持平行處理、高效利用緩存，適用於分佈式和大數據集的計算。
- **高度靈活性（High Flexibility）**:提供自定義目標函數和評估指標。
- **特徵重要性分析（Feature Importance Analysis）**:對每個變量的相對重要性進行排名。
- **樹修剪（Tree pruning）**: 實現了樹修剪分割點處理。



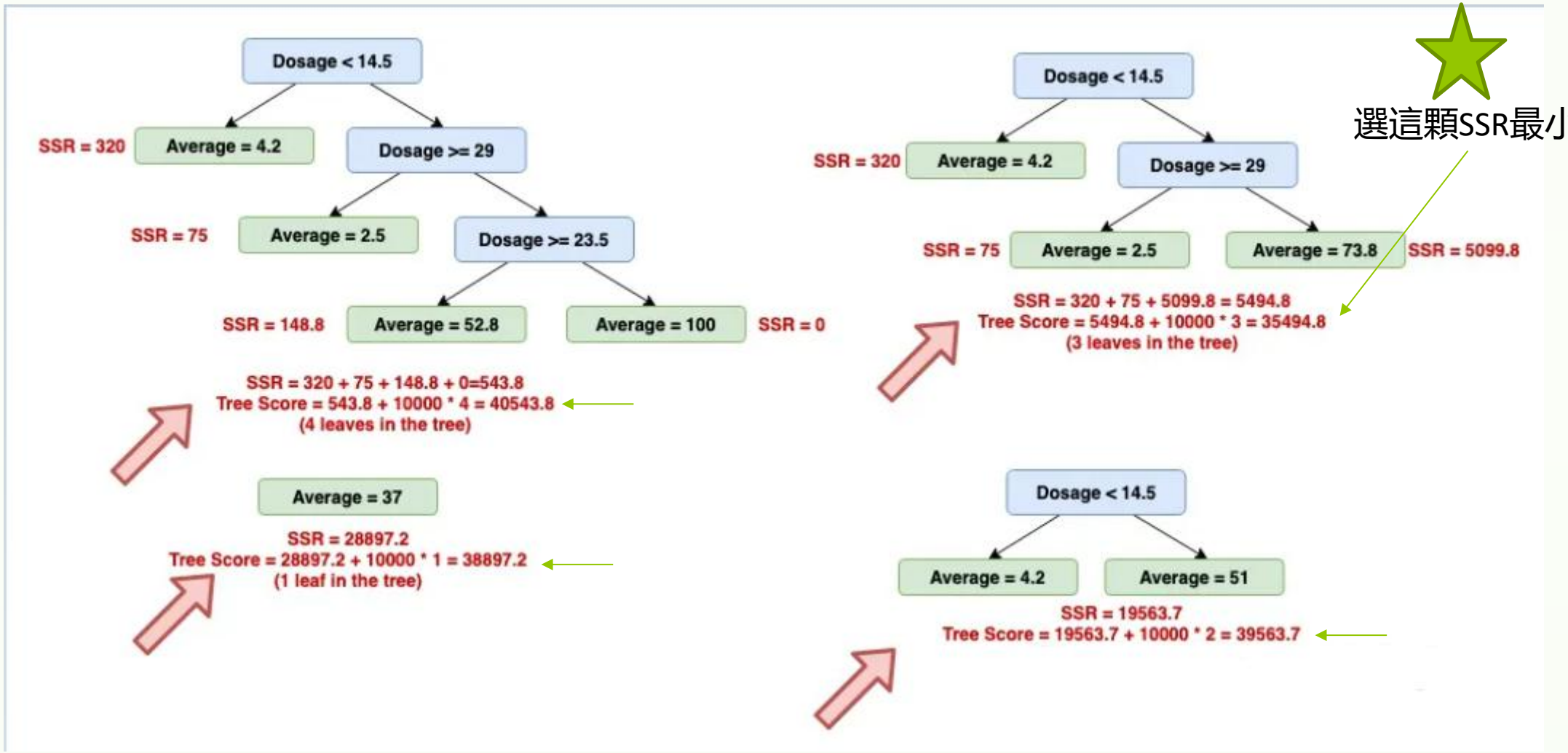
SSR計算葉與樹的誤差平方合

- **計算sum of the squared residuals (SSR) 誤差平方合, 每片葉子, 再加總整個樹。**



https://zhuanlan.zhihu.com/p/584124751

同一顆對不同修剪結構選最SSR最小



AI 實作之參數與資料集的校調

- **機器學習方法(AI的子集), 屬資料導向(Data driven)**
- **以非線性模形大量資料的樣版(Pattern)預測股價走勢**
- **參數與資料集改變會改變預測結果;**
- **資料集校調**
  - **LSTM: 採用前2日收盤價大於隔日收盤價為上漲; 反之為下跌，準確率達78%。**
  - **XGBoost: 採用前2日收盤價與當日收盤價的百分比，準確率達78%。**
- **輸入變數**
  - **統計獨立變數與相依變數因果關係分析**
  - **技術指標、財報指標、總經指標**
  - **國際期刊輸入變數與股價或報酬率的預測效果**

參考文獻

1. Han, Y. C., Kim, J., & Enke, D. (2023). A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost [Article]. *Expert Systems with Applications, 211, 10, Article 118581.* https://doi.org/10.1016/j.eswa.2022.118581

2. Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process [Article]. *Expert Systems with Applications, 186, 21, Article 115716.* https://doi.org/10.1016/j.eswa.2021.115716

3. Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & Shahab, S. (2020). Deep Learning for Stock Market Prediction [Article]. *Entropy, 22(8), 23, Article 840.* https://doi.org/10.3390/e22080840

4. Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers [Article]. *North American Journal of Economics and Finance, 47, 552-567.* https://doi.org/10.1016/j.najef.2018.06.013

5. Ampomah, E. K., Qin, Z. G., & Nyame, G. (2020). Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement [Article]. *Information, 11(6), 21, Article 332.* *https://doi.org/10.3390/info11060332*

簡報排版非正式參考文獻格式，敬請見諒。

章節到此結束，有任何問題歡迎提出來討論！