1.Introduction

我們做的是股票預測系統，股票作為理財、學術、多樣性等各方面來說都是一個天然的理想資料庫，他具備變化性合理性以及一定的不確定性，昨為預測模型的建立和推算算是一個經典的案例。

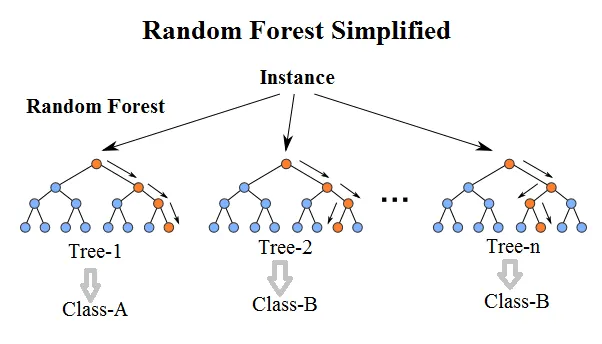
以學生的觀點來說，參與這樣的嘗試可以對我們未來的求職、對課堂學習知識的實踐與現實實際問題的解答、和學習股票理財的知識都是一項很好的練習學習機會，在求職方面現在的資料工程師可以說是一項很好的出路，學習處理資料整合資料是我們能夠練習的一次體驗，還有許多演算法和資料結構握們大多都只在課堂上講解過，就算有練習多半也是淺嚐而止。只有經過實際數據的練習才是真的有學以致用，最後理財幾乎是每個人生中不可避免的一部分，而股票投資也是投資理材中很重要的部分。以dss來說，整理資料，建立模型，訓練模型，判斷決策是我們在這次題目中使用到的部分。綜合以上，我們選擇了股票預測這樣的主題。

我們希望我們的模型藉由過去的紀錄判斷出某支股票未來的漲跌，實際上我們也發現了在網路上，論文中早已存在許多的預測方式，但是因為實際股票充滿了許多的不確定性，除非你有內線消息，不然目前最好的辦法依舊是利用經驗法則來進行預測，所以我們沒有把目標設定的太高，精準的預測是很難的，我們只打算做出一個能簡單預測漲跌的程式，藉由演算法去判斷未來的股票變化，不過這依舊很難，因為單看數據很難讓程式去感受到國際大勢變化的轉折點，以目的來說我們只能希望盡可能讓他符合單純的實際漲跌，但是太多的需要判斷因素導致希望達到的是漲是跌這件事情，都相當困難。

2.Related works

在技術上我們使用了從網頁抓取資料的資料整理網路爬蟲，演算法和介面程式是我們用到的部分。

相關技術1:隨機森林



隨機森林是一個強大的機器學習方法，其基本概念是建立多個決策樹並將它們整合為一個預測模型。

### 決策樹（Decision Trees）：

* 隨機森林的基礎是決策樹。決策樹是一種樹狀結構，用於將資料分成不同的類別或進行預測。
* 每個決策樹都根據資料特徵進行分割，形成多個節點和分支，最終形成葉子節點，代表預測結果。

### 隨機性：

* 隨機森林引入隨機性來提高模型的多樣性和準確性。
* 通過對訓練數據進行隨機抽樣（bootstrap抽樣）來建立不同的數據集，並且在每個決策樹的構建過程中，隨機選擇特徵來進行節點的分割。

### 整合預測：

* 每個決策樹為資料提供預測結果。在隨機森林中，將這些決策樹的預測結果進行整合，例如對於分類任務，通過多數投票決定最終的類別；對於回歸任務，則取平均值作為最終預測。

### 優點：

* 隨機森林有效避免了過擬合問題，因為它基於多個決策樹的整合，並且使用隨機抽樣和隨機特徵選擇。
* 能夠處理大量資料和高維度特徵，同時對缺失值和雜訊具有良好的魯棒性。
* 可以評估特徵的重要性，幫助了解哪些特徵對於模型的預測起到更大的作用。

步驟

1. 定義大小為n的隨機樣本（這裡指的是用bagging方法），就是從資料集中隨機選取n個資料，取完後放回。
2. 從選取的n個資料中，訓練出決策樹。對每一個節點：  
   a. 隨機選取d個特徵  
   b. 使用特徵分割該節點
3. 重複k次步驟1~步驟2
4. 匯總所有決策樹的預測，以多數決的方式，來決定分類結果。

相關技術2:CNN-resnet

CNN是像發展已久的神經網路學習模型，相較其他模型更加成熟。其中，我們又加入了residual path來修復他可能會梯度消失的問題

相關技術3:網路爬蟲

網路爬蟲是一種自動化程式，用於從互聯網上的網頁中檢索資訊。它們以自動方式訪問網站，獲取網頁數據並將其存儲或進行進一步處理。

### 工作原理：

* HTTP 請求： 爬蟲會向目標網站發送 HTTP 請求，請求特定網頁的內容。
* HTML 解析： 爬蟲獲得響應後，會解析 HTML 代碼以提取有用的資訊，如文本、圖像、連結等。
* 遞歸抓取： 爬蟲可以根據初始頁面上的連結，遞迴地抓取其他頁面，構建整個網站的檔案或索引。

### 用途：

* 搜集資訊： 爬蟲常用於搜集大量資訊，例如搜索引擎的索引建立、資訊檢索和分析等。
* 監控和更新： 可用於定期監控網站的變化，例如價格變動、新聞更新等。
* 數據分析： 爬蟲可以收集大量資料用於分析，例如市場趨勢、社交媒體數據等。

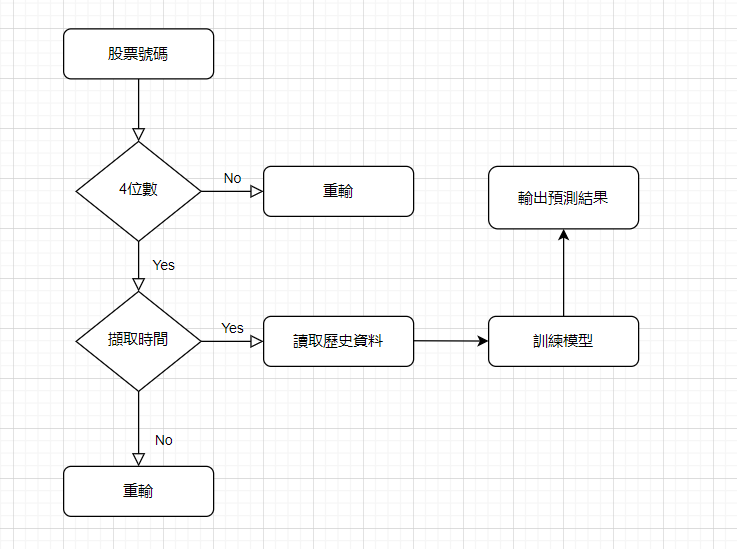
步驟

1. 向目標網站請求特定網址（URL）的內容

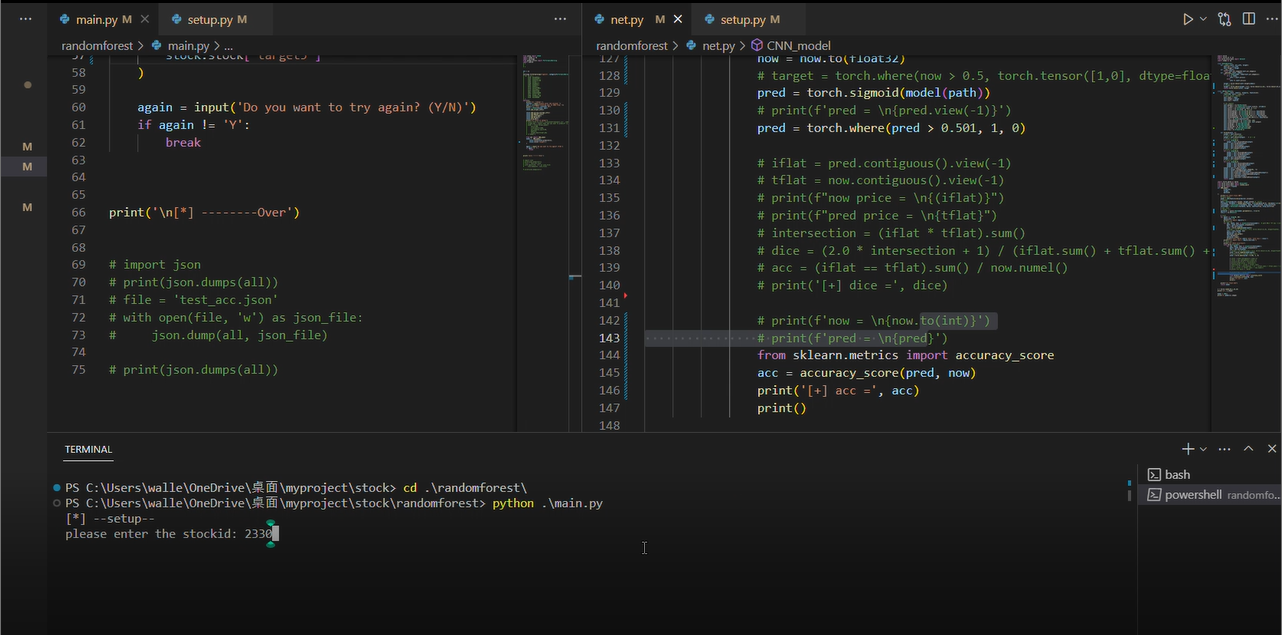
2. 將 HTML 文件做「解析」並「取出」所需的資料

3. 將取出的資料儲存在 CSV 檔案、Excel 表或是資料庫當中

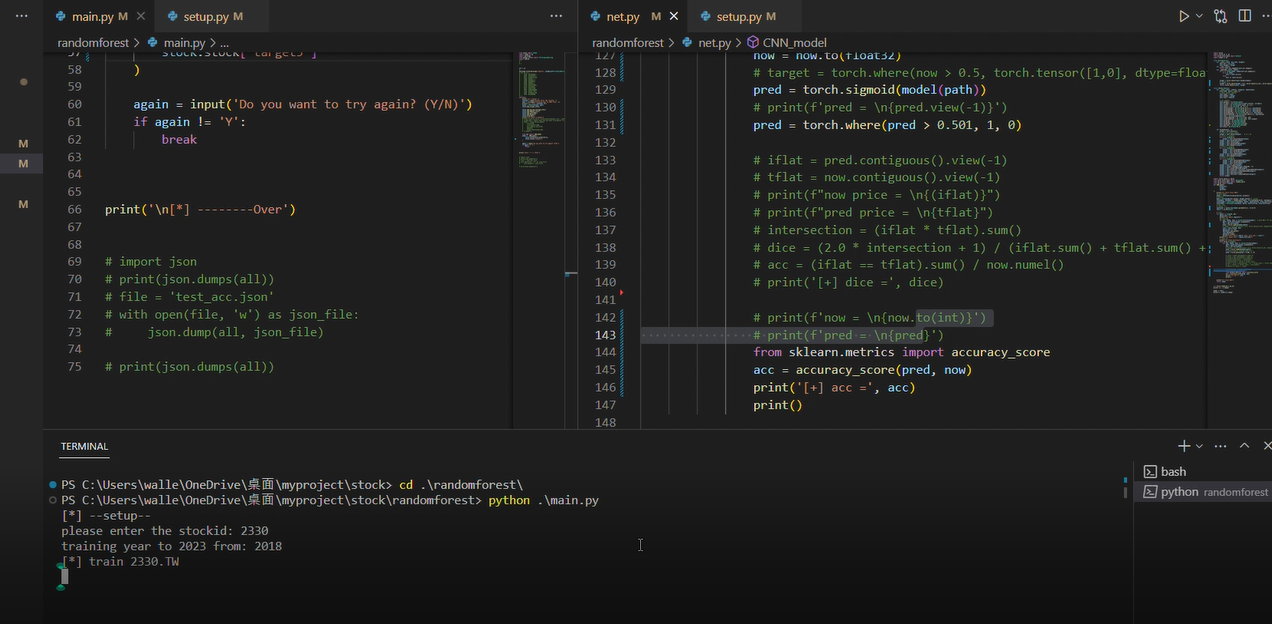
3.The proposed approach



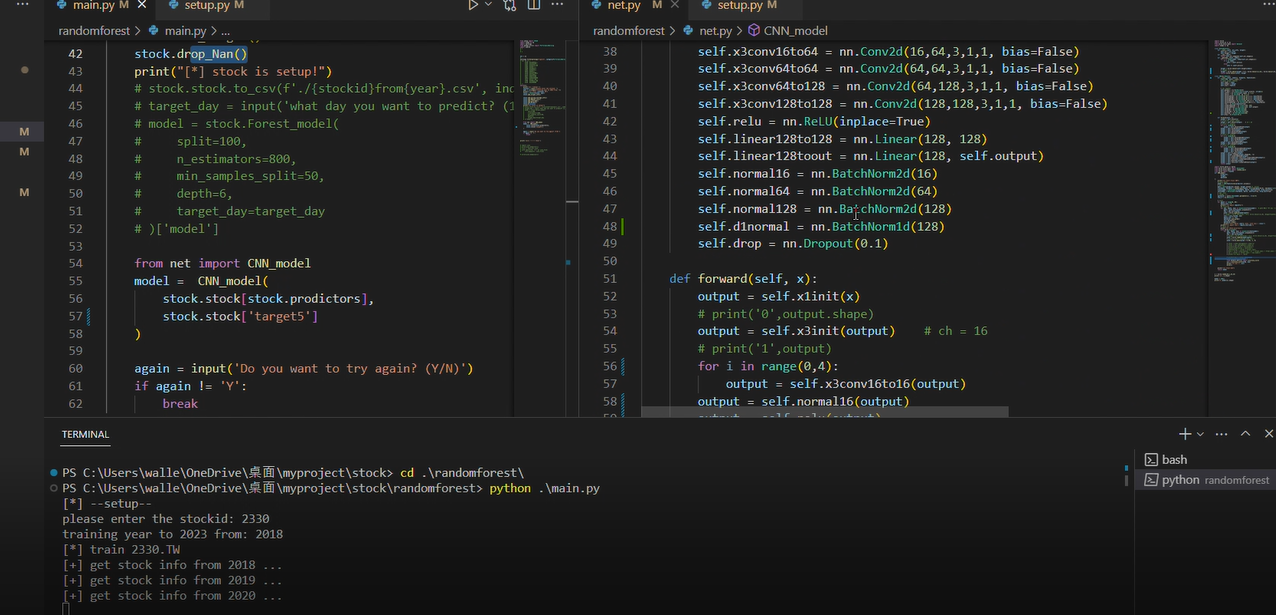
1.輸入股票



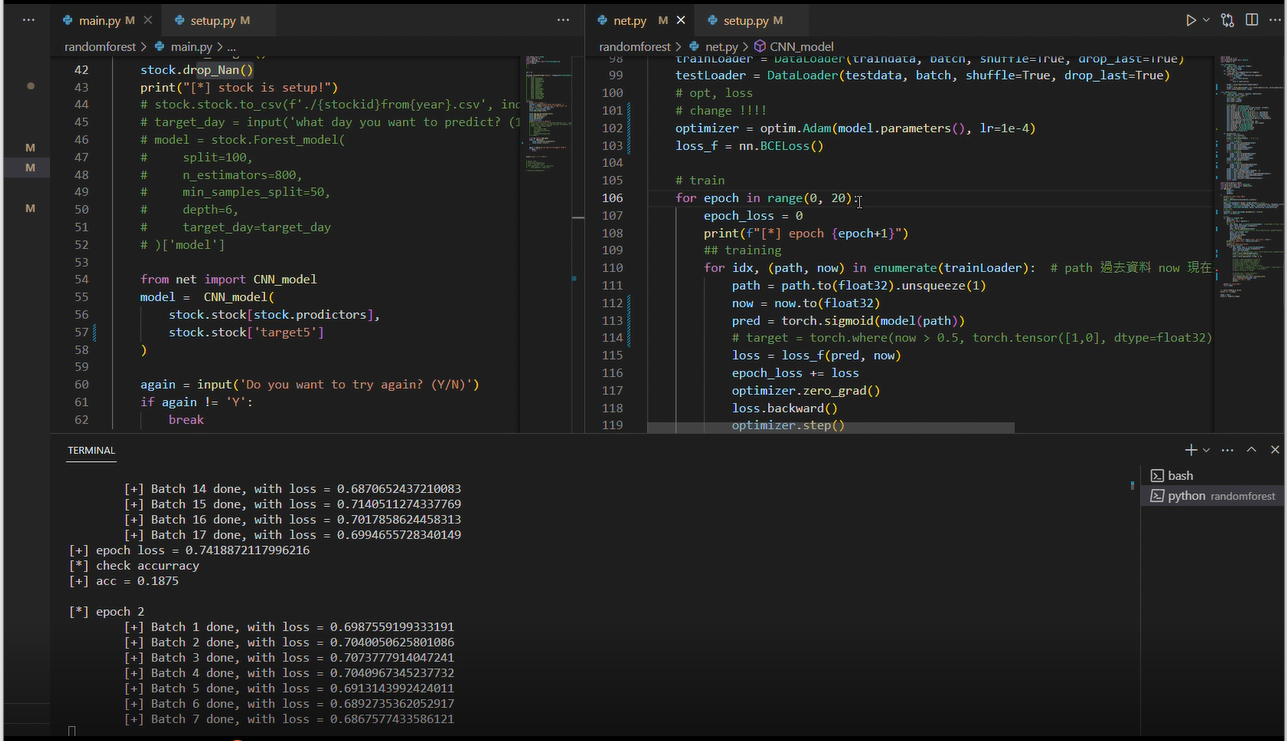
2.輸入時間



3.網路爬蟲抓資料



4.訓練模型



4.Experimental results

5.Conclusio

單看結論可能和當初所設想的結果有一定的出入，例如想要出現的結果和預測的準確性等等，尤其是近年台股的漲幅變化幅度之劇烈，更是加大了預測和訓練的難度，這可能也是股票預測一直難以系統化原因，不過總體而言，還是讓我們獲益良多，是一個很好的實踐經驗。

6.Appendix

組員：許少華 法丞駿 翁紹期 邱閔聖 沈韋仲

資工三B 409411732 資工三B 410411432 資工三B 410411192 資工三B 410410806 資工3B 410410384資工3B 410411259

參考資料網站：

<https://www.webscrapingpro.tw/what-is-web-scraping/>

<https://medium.com/chung-yi/ml入門-十七-隨機森林-random-forest-6afc24871857>

https://medium.com/ai%E8%82%A1%E4%BB%94/%E5%AD%B8%E6%9C%83%E7%94%A8%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E9%A0%90%E6%B8%AC%E8%82%A1%E5%83%B9-%E5%AE%8C%E6%95%B4%E6%B5%81%E7%A8%8B%E6%95%99%E5%AD%B8%E8%88%87%E5%AF%A6%E4%BD%9C-b057e7343ca4

https://blog.csdn.net/weixin\_43704656/article/details/120921289

https://ithelp.ithome.com.tw/articles/10289229

https://pypi.org/project/yfinance/

https://yhhuang1966.blogspot.com/2022/09/python-yfinance.html

https://beautiful-soup-4.readthedocs.io/en/latest/

https://twstock.readthedocs.io/zh-tw/latest/index.html

https://www.youtube.com/watch?v=2S1dgHpqCdk&list=PLhhyoLH6IjfxeoooqP9rhU3HJIAVAJ3Vz

https://www.stockfeel.com.tw/

https://www.tejwin.com/insight/%e3%80%90%e8%b3%87%e6%96%99%e7%a7%91%e5%ad%b8%e3%80%91%e9%9a%a8%e6%a9%9f%e6%a3%ae%e6%9e%97%e6%bc%94%e7%ae%97%e6%b3%95%e9%81%b8%e8%82%a1%e7%ad%96%e7%95%a5/

https://scikit-learn.org/stable/index.html