





# AI Cup

**Introduction to Data Science**

# CONTENT

- 
- 
- 01** COMPETITIONS
  - 02** **COMPETITION 1**
  - 03** DATA SETS
  - 04** ANALYSIS PERFORMED
  - 05** COMPETITION RESULTS
  - 06** **COMPETITION 2**
  - 07** DATA SETS
  - 08** ANALYSIS PERFORMED
  - 09** COMPETITION RESULTS

# COMPETITIONS



根據區域微氣候資料  
預測發電量競賽



玉山人工智慧公開挑戰賽  
RAG與LLM在金融問答的應用





# COMPETITION 1

根據區域微氣候資料  
預測發電量競賽

# DATA SETS

- 資料包含大約2000天的數據，提供17個太陽能監測地點的太陽能板設備附近之微氣候數據
- 7個特徵：
  - 地點代號LocationCode(1-17)
  - 時間DateTime(Y/M/D/hour/minute/second)
  - 風速WindSpeed(m/s)
  - 大氣壓力Pressure(hPa)
  - 溫度Temperature(°C)
  - 濕度Humidity(%)
  - 亮度Sunlight(Lux)
- 1個標籤：太陽能板每分鐘平均發電量Power(mW)
- 訓練資料的時間序列每分鐘一列，但有時會有資料缺失
- 預測某一天9:00(含)之後的每一筆(每10分鐘)的平均發電量(mW)，持續預測直至當天16:59(含)



# ANALYSIS PERFORMED

## Preprocess

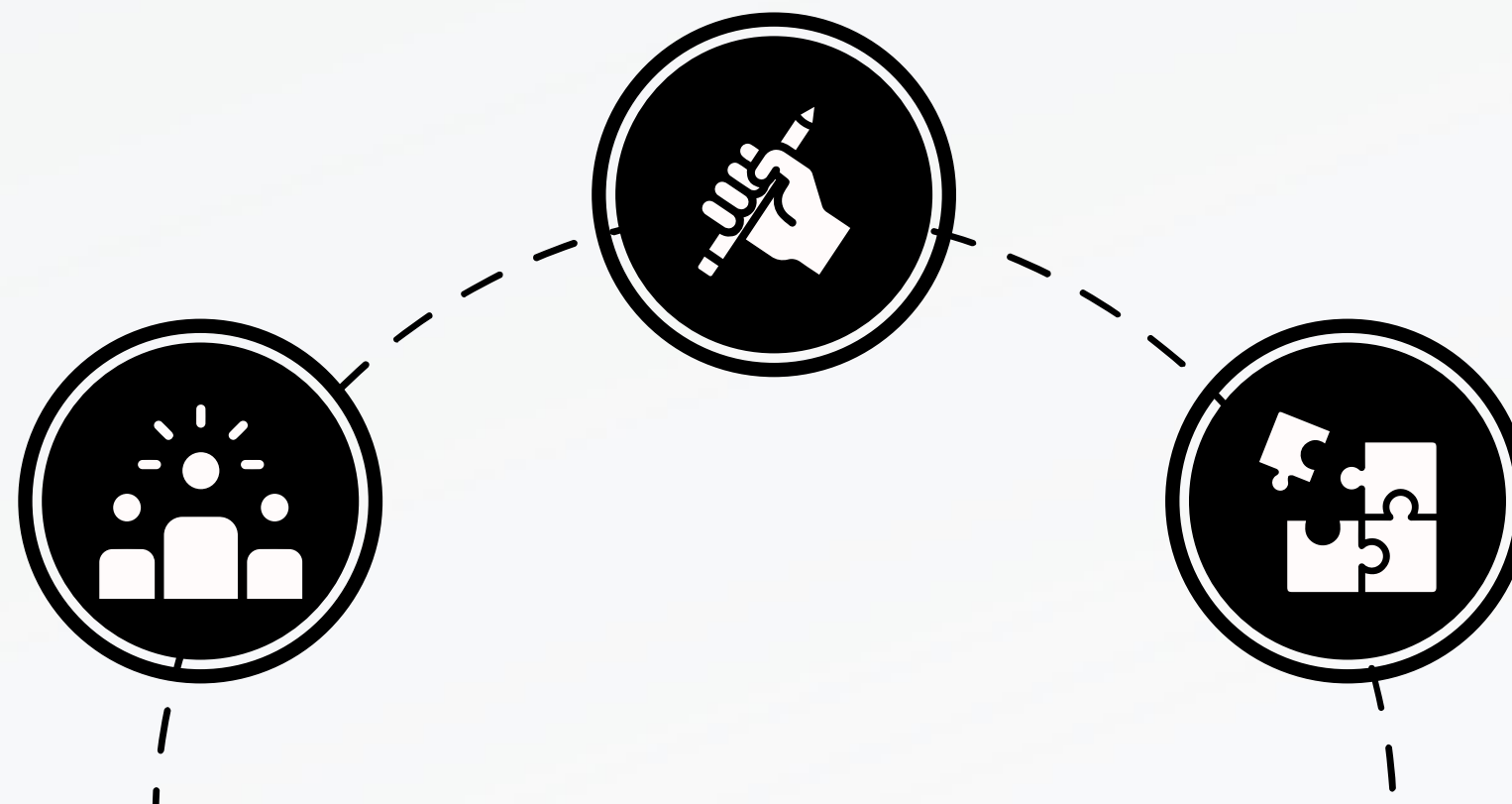
將日期與時間拆分為  
Year、Month、Day、  
Hour、Minutes、Second  
拆分訓練集、測試集

## Train Model

simple regression  
multiple regression  
support vector regression  
多種Scikit-learn Regression

## Predict

使用訓練好的模型進行預測  
計算均方誤差MSE  
MSE越小越好





# MODEL

XGBoost是一種基於梯度提升演算法的高效機器學習工具，廣泛應用於迴歸和分類任務，而XGBoost Regressor專門用於解決迴歸問題，適合預測連續數值，例如房價、銷售額、氣象指標等

重要參數：

max\_depth：控制過擬合(3~10為佳)

learning\_rate：迭代的步長(0.1左右)

n\_estimator：最大的迭代次數



# XGBoost Regressor




# MODEL

我們嘗試過以下Model：

Simple、Multiple、Support Vector、  
Ridge、Lasso、KNN、Decision Tree、  
AdaBoost、Gradient Boosting、  
Random Forest、ANN、RNN、LSTM、GRU

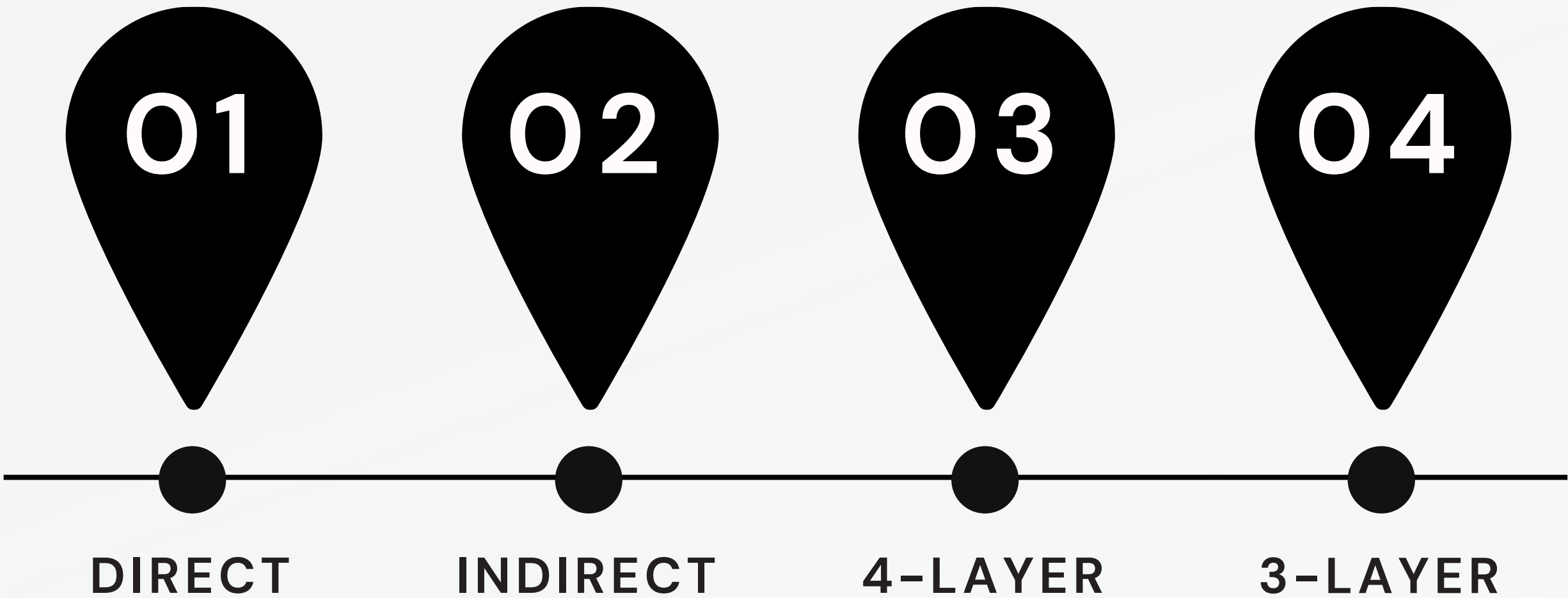
但有些Model不但執行速度慢，正確率也低，因此我們在考量正確率及執行速度等因素之後選擇XGBoost



# OTHER MODELS



# PREDICT METHOD



直接用時間地點  
同時預測其他值  
與發電量

先用時間地點預  
測其他值，再由  
其他值預測發電  
量

採取4層方式，  
每層皆由前面的  
層預測

發現風速與氣壓  
為雜訊，因此將  
非必要的第一層  
刪除

BASIC	時間、地點
LAYER 1	風速、氣壓
LAYER 2	溫度、濕度
LAYER 3	亮度
TARGET	發電量

# COMPETITION RESULTS

隊伍名稱

**TEAM\_6373**

934

參賽隊伍

Public Leaderboard			Private Leaderboard		
#	隊伍名稱	成員	提交次數	分數	上傳時間
34	TEAM_6373	4	28	692941.36	11/28/2024 11:35:09 PM
Public Leaderboard			Private Leaderboard		
#	隊伍名稱	成員	提交次數	分數	上傳時間
35	TEAM_6373	4	28	768296.3	11/28/2024 11:35:09 PM



# COMPETITION 2

玉山人工智慧公開挑戰賽  
RAG與LLM在金融問答的應用

# DATA SETS

- `questions_example.json`為範例題目
- `ground_truths_example.json`為範例題目的答案
- `pid_map_content.json`為玉山銀行官方網站上的常見問題之資料集
- `insurance`為玉山銀行代銷的保險產品之保單條款的PDF文件資料
- `finance`為公開資訊觀測站上的上市公司財務報告的PDF文件資料
- 依據題目的問題分析答案能在哪個文件的內容中找到

# ANALYSIS PERFORMED

- 處理中文斷詞
- 檢索金融相關關鍵字作為字典
- 將資料中的空白、換行、標點符號替換掉

Preprocess

使用LMIR演算法進行資料檢索：

- 將每篇文檔及查詢語句進行分詞
- 計算資料庫當中的詞頻
- 計算每篇文檔的長度和詞頻
- 根據查詢語句進行檢索，找到最佳匹配的文本並回傳檔名

Data Retrieval

計算匹配比例及每個類別的準確度

Matching



# PREPROCESS

- 有嘗試過將PDF中的**表格**讀取出來串接在PDF的文字後面，但發現正確率沒有顯著的提升，因此我們決定不讀取表格
- 從PDF中讀取**圖片**太過耗時，且圖片大部分是公司LOGO，是對分析沒有幫助的雜訊，因此也不讀取圖片
- **insurance**跟**finance**兩個類別在讀取PDF時需要清洗大量斷詞，但同樣的方法不適用於FAQ類別的純文字檔，正確率會下降

# 檢索金融相關關鍵字

## INTRODUCTION



jieba中的dictionary能夠將常見詞彙被正確的分詞出來  
我們需要增強在金融方面的應用（訓練資料很多是契約書和有關一些統計的資料）  
加入行業術語應該有助於更正確地找到解答

## EXAMPLE: dict.txt



臺灣科技大學  
自然語言處理  
人工智慧

```
jieba.set_dictionary('dict.txt')
```

使用前：我/在/臺灣/科技/大學/學習/自然/語言/處理/與/人工/智慧

使用後：我/在/臺灣科技大學/學習/自然語言處理/與/人工智慧

## METHOD



利用Chat GPT這個特別強的LLM去幫我們找出哪些詞彙應該要被放入dictionary，將題目敘述、所有的文件選項的文字檔與Ground Truth放入同一個文件中，讓GPT閱讀過後選出他認為適合作為關鍵字的字串



# MODEL

LMIR是基於語言模型的資料檢索方法，它藉由計算某個查詢由特定文件生成的機率來排序文件。LMIR最常見的應用是從文檔集中找出最匹配查詢的文檔。

Dirichlet平滑參數  $\mu$ ：

$\mu$  大適合長文件、 $\mu$  小適合短文件

Dirichlet平滑適合大多數應用場景，因為它在處理長短文件方面有自適應特性，適合需要高準確性的檢索



# LMIR Retrieve





# MODEL

我們嘗試過以下Model：

BM25、DSMM、BERT、TFIDF、ANN、  
Doc2Vec、LSI、LDA、NMF

並搭配斷詞、資料清洗及關鍵字，找到在正確率及執行速度上的最佳組合為LMIR



# OTHER MODELS

# COMPETITION RESULTS

隊伍名稱

**TEAM\_6372**

487

參賽隊伍

## Public Leaderboard

## Private Leaderboard

#	隊伍名稱	成員	提交次數	分數	上傳時間
64	TEAM_6372	4	3	0.87472	11/9/2024 2:30:40 PM





## 玉山銀行人工智慧公開挑戰賽

### Certificate of Participation

王家宏

於 2024 年 11 月 9 日參加  
玉山銀行人工智慧公開挑戰賽 2024 冬季賽  
並獲得 60th/ 218 名之成績  
特頒予此參賽證明，以茲紀念。

This certificate is rewarded to

Wang Jia Hong

in recognition of his/her participation in  
E.SUN AI Open Competition Winter 2024  
and placed 60th out of 218 participants  
from 9<sup>th</sup> November, 2024.

 玉山金控 玉山銀行

**THANK'S FOR  
WATCHING**

