**建立預測模型 與 特徵選擇**

**特徵 :**

**目標值** :

ABS(明日收盤價 – 今日收盤價 / 今日收盤價)\*100 > 0.8 為 TRUE

並確認抓到每一個行情日。

# 可以由交易員提供行情日做為目標值

**特徵** :

當沖成交金額、 大盤成交金額、 VIX振幅、 OTC成交量變動、 Gamma、 IV 、 STW未沖銷契約數、 TX未沖銷契約數、 兌美元匯率、 STW OI \* STW close \* 100 \* 兌美元匯率(摩台未平倉金額) – TX OI\* TX close \* 200(台指未平倉金額) (考慮市場中做ALPHA的交易者有多少)

**進行方式 :**

使用機器學習(RandomForest, Logistic, XGBoost)建立 預測模型

同時

1. 根據小組建議，改良特徵

2. Model Ensemble 模型融合

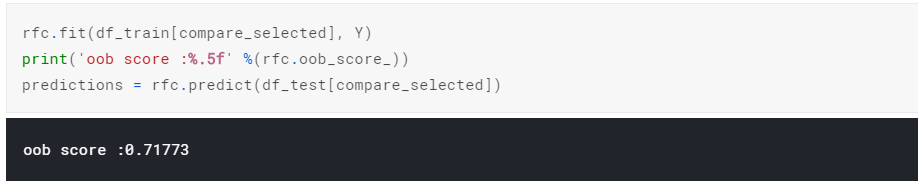
**特徵選取方式 :**

懲罰化(Lasso, Ridge), test檢定, 決策樹熵增益, Python套件sklearn 中提供的方法

優先選取各個方法中 重要性高 共同前幾名 特徵

排除各個方法中顯示 重要性低 共同後幾名 特徵

排除相關性高的特徵



oob score 準確率為71.773% (test 某次決策樹中沒有被選取的數據)

**問題 :**

目標值改為 (明日收 – 夜盤收) ? 可由交易員自行給定目標值

* 若抓早期資料一併訓練，模型訓練完後，早就不適用於現在市場
* 拿近期資料訓練，失去優勢? 實驗後也發現準確率沒有提升

是否需要考慮 期貨轉倉 價格缺口 ? A: 需排除 7, 8, 9月換月合約價格缺口

目前所找的特徵在 行情日 與 平常日，資料特性過於相似

**總結 :**

1. 實際test資料預測效果差

2. 沒有考慮 波動群聚性 與 時間

**這週目標 :**

使用 “時間序列” 相關方法來創建模型(越近期資料越重要)

**考慮問題 :**

季節性(傅立葉轉換 STL分解)

如何判定 落後期數選取(AIC, BIC)

單根檢定(資料定態分析)

檢定異常點(去年被動元件、台指閃崩)

**現在方向 :**

1. 建構神經網路模型 LSTM, GRU (與特徵結合)

需要較大運算量

2. 建立一套尋找 重要特徵 的 SOP

3. 需有一套方法告知交易員 特徵 如何使用