模型訓練及選用問題

如果要達成「能夠預測指定的symbol在90天後，是否有成長10%」的目標，會選用的模型及訓練方式，文字提供選用的模型及原因

為了達成「能夠預測指定的 symbol 在 90 天後是否成長 10%」的目標，本研究將此問題設定為一個帶時間約束的二元分類任務。每一個樣本代表在某一日 t 的個股狀態，其標籤由 90 天後的價格變化定義：若未來收盤價 P_{t+90} 與當日收盤價 P_t 的比值減一大於或等於 0.1，則標籤為 1，否則為 0。對無法完整觀察到未來 90 天價格的樣本予以剔除。為防止資訊外洩，所有特徵均僅使用當日及之前能夠取得的資料，季度財務類指標以報告對齊日為界。

模型訓練採用時間序列切分方式，確保訓練集、驗證集與測試集在時間上嚴格分離，避免傳統隨機分割帶來的未來資訊穿越。以 2021−2023 年為訓練集、2024 年為驗證集、2025 年年初資料為測試集。模型評估指標選用 ROC–AUC 與 PR–AUC，並輔以 Precision、Recall 與 F1；在業務層面，亦會依預測機率排序建立 Top–K 組合進行 90 天持有回測，以驗證策略收益與穩定性。

在模型選擇上，邏輯迴歸被用作基線模型，用以建立可解釋的機率框架。邏輯迴歸透過 sigmoid 函數將線性組合的特徵映射為上漲機率，能夠直觀展示各特徵對目標的正負影響，但其線性假設無法有效捕捉金融資料中高度非線性與交互性訊號。因此，在主力模型上選用 LightGBM。LightGBM 屬於梯度提升決策樹家族，採用葉節點優先的生長策略與基於直方圖的梯度計算，能高效率地學習複雜的非線性關係，自動處理缺失值與特徵尺度問題，並提供基於增益或 SHAP 值的特徵重要度分析。在大規模結構化表格資料上，它兼具高精度與高速度。訓練過程中控制樹深、葉數及學習率，結合 L2 正則、樣本及特徵子抽樣（bagging_fraction、feature_fraction）以抑制過擬合，同時設定類別權重 scale_pos_weight 處理正負樣本比例失衡。

在訓練中如何從現有的資料集提取出關鍵影響欄位

三個層次進行：模型內部的重要性分析 → 模型外部的驗證 → 經濟意涵的詮釋。

首先是模型內部的特徵重要性提取。以 LightGBM 這類樹模型為例，演算法在每次節點分裂時都會計算一個「資訊增益」（Gain），代表該特徵能讓損失函數下降多少。模型訓練完成後，可根據各特徵在所有樹中的累計增益、分裂次數或樣本覆蓋量，得到全域的重要性排序。這一步能快速指出哪些變數在模型中最常被使用、貢獻最大。例如，近二十日報酬率、成交量變化率、毛利率、ROE、自由現金流成長率等指標，往往在金融預測任務中名列前茅。

第二層是模型外部的重要性驗證。單靠增益排序可能會受模型結構偏好影響，因此需採用更穩健的「置換重要性」（Permutation Importance）方法：將某個特徵的數值隨機打亂，再重新計算模型的預測精度，若精度下降幅度大，表示該特徵越關鍵。此外，也可運用 SHAP（SHapley Additive exPlanations）分析每個樣本在各特徵上的邊際貢獻。SHAP 值基於博弈

論原理，能同時呈現特徵對模型輸出的正負方向與貢獻強度，不僅能在全域層面觀察哪些特徵整體提升上漲機率，也能在單一樣本層面解釋模型的個別判斷。

最後是結果的經濟與統計詮釋。重要性排序僅顯示數值層面的權重，仍需結合金融邏輯檢驗其合理性。例如，若模型顯示「營收成長率」、「營運現金流成長率」及「價格變動率」為主要正向特徵，這符合市場中成長股上漲的邏輯；若「負債比」、「波動率」或「估值倍數」呈現負向貢獻，則意味市場傾向避開高槓桿或高估值標的。此步驟能幫助確認模型捕捉到的確為真實經濟訊號，而非資料雜訊或偶然現象。

如何利用目前已有的資料集欄位，推論出更有效的新資料欄位

可從三個層面設計新特徵：時間維度延伸、交互與比例關係、以及風險與動量因子衍生。

第一層：時間維度延伸（Temporal Features）
從現有的日線或季度資料中，計算不同時間窗口的統計特徵。例如：

報酬動能特徵：利用過去 5、20、60 日的價格變化率、對數報酬率或累積報酬率，以反映趨勢強度。
波動率特徵：計算近期期間報酬率的標準差或高低價區間差（high–low range），以刻畫風險水準。
量價變化特徵：如成交量相對 20 日均量的比例、成交金額變化率、價量背離指標（例如價漲量縮）。
移動平均與乖離率：例如 close / MA20 – 1，用以反映價格相對均線的偏離程度。
財報期接近度：利用 period_start 與 period_end 計算「距期末天數」或「是否處於財報揭露窗口」，以捕捉財報效應。

第二層：交互、比例與結構關係（Cross & Ratio Features）
透過構造特徵之間的比值或交互項，揭示潛在的結構性關係。

估值調整特徵：例如 PE × 成長率（PEG）、PB / ROE（相對估值指標）。
償債與營運效率比：以負債比 / 現金流成長、營收成長 / 資產成長衡量資本使用效率。
量價交互項：價格變化 × 成交量變化，用來區分趨勢放量與縮量情境。
非線性轉換：對偏態分布的特徵（如市盈率、成交量）取對數或平方根，緩解極端值影響。
特徵組合編碼：將產業、月份等類別特徵與動能或估值指標組合，例如「產業 × ROE」或「月份 × 波動率」。

第三層：風險、品質與動能因子衍生（Factor-style Features）
結合現有的財務與市場特徵，推導出類似量化因子的綜合指標。

成長性因子：營收成長率、淨利成長率、自由現金流成長率等的滾動平均或穩定性。
獲利品質因子：ROE、ROA、毛利率、營運現金流 / 淨利等，用以衡量獲利的持續性。
槓桿與風險因子：負債率、現金流波動率、Beta、波動率 / 報酬率比。
市場動能與反轉因子：短期動能（520 日）與中期動能（60120 日）方向相反時的反轉訊號。
財報窗口因子：結合 calendarYear + period 生成季度虛擬變數或公告日前後窗口變數，以捕捉季節性效應。