JPX Tokyo Stock Exchange Prediction

Transformer 架構與特徵工程深度解析

Team9 | National Central University

團隊成員

114552022 張敬慧　114552029 馬廣嫻

目錄

摘要 2
1. 競賽背景與評分標準 3
2. 數據來源與預處理 4
3. 特徵工程 — 22 個特徵的設計 7
4. 數據的證據：特徵統計分析 8
5. 量級差異問題與 L2 正則化 16
6. 訓練策略：穩健的損失函數與超參數權衡 17
7. 模型架構 — 2 層 Transformer 17
8. 模型比較：Transformer vs LSTM vs MLP vs LSTM-Transformer 18
9. 專案成果總覽 23
10. 討論與未來工作 23
11. 結論 24

Source code：∞ DL_lab3_team9

摘要

本報告描述了 Team9 在 JPX Tokyo Stock Exchange Prediction 競賽中開發的股票收益預測系統。我們採用 Transformer 深度學習架構，結合精心設計的 22 個特徵（其中 68% 為時間依賴特徵），成功預測東京股票交易所的股票相對收益排名。系統使用最近 100 個交易日的歷史數據進行訓練，並通過 L2 正則化解決特徵量級差異問題。最終模型在競賽中取得了 0.390 的 Sharpe Ratio，證明了深度學習在金融時間序列預測領域的應用潛力。

1. 競賽背景與評分標準

1.1 競賽要求

JPX Tokyo Stock Exchange Prediction 競賽提供從 2017 年至今約 2000 支股票的每日交易數據，要求參賽者預測未來 2 個交易日的相對收益排名（Relative Return Ranking）。

1.2 預測目標

預測每支股票在下一個交易日的相對收益，並根據預測結果進行排名，以支持市場中性策略（Market Neutral Strategy）的構建。

1.3 交易策略

基於預測排名，做多（Long）前 200 名股票，做空（Short）後 200 名股票，構建市場中性投資組合。

1.4 評分標準

競賽使用 Sharpe Ratio 作為評估指標：

Sharpe Ratio = E[Return] / Volatility

該指標衡量投資組合的風險調整後收益，數值越高表示在相同風險下獲得更高的收益。

2. 數據來源與預處理

2.1 五大數據來源

- **Stock Prices**（股票價格）：包含開盤價、收盤價、最高價、最低價、成交量等基本行情數據
- **Financials**（財務數據）：財報相關信息
- **Stock List**（股票清單）：股票基本資料
- **Options & Trades**（衍生品交易）：期權和交易數據
- **Secondary Prices**（次級股票）：輔助價格信息

2.2 關鍵決策：使用 AdjustedClose

2.2.1 股價斷層問題

在金融市場中，股票因分割（Stock Split）、合併或發放股利（除權息）時，其收盤價（Close）會出現非市場因素的劇烈跳動。若直接使用未經處理的收盤價，模型會將這些價格斷層誤判為市場的真實暴漲或崩盤，從而學習到錯誤的規律。

2.2.2 解決方案

使用調整後收盤價（AdjustedClose）消除除權除息造成的虛假跳幅。計算公式為：

AdjustedClose = Close × CumulativeAdjustmentFactor

其中 CumulativeAdjustmentFactor 為調整因子的累積乘積，由歷史到現在倒序計算。

2.2.3 影響

使用 AdjustedClose 後，成功消除虛假信號，提升了 22 個特徵的品質，使模型能夠專注於真實的市場動態。

2.3 數據預處理流程

在特徵工程之前，我們對原始數據進行了嚴格的清洗和預處理，確保數據質量。完整的預處理流程包括缺失值處理、異常值過濾、數據標準化等關鍵步驟。

2.3.1 缺失值（NaN）處理

金融數據常見缺失值問題，主要來源包括：股票停牌或暫停交易、新上市股票缺乏歷史數據、數據源錯誤或傳輸問題、計算特徵時窗口不足（如新股的 MA60）。

處理策略：分層處理法

策略 **1**：基礎價格數據 - 前向填充（**Forward Fill**）

- 適用欄位：Close, Open, High, Low, Volume
- 方法：使用前一個交易日的數據填充
- 原理：股票價格具有連續性，前一日收盤價是當日開盤價的良好估計
- 程式碼：df.fillna(method='ffill')

策略 **2**：計算特徵 - 零填充（**Zero Fill**）

- 適用欄位：RSI, MACD, Return_1d, Return_Mean_5d 等技術指標
- 方法：計算結果為 NaN 時填充為 0
- 原理：收益率為 0 表示無變化（中性值）；RSI 為 0 表示超賣，MACD 為 0 表示無趨勢，都是有意義的狀態

- 優勢：避免引入偏差（使用均值填充可能導致模型過度依賴某些值）

策略 **3**：移動平均 - 回退策略（**Fallback Strategy**）

- 如果 MA60 缺失 → 使用 MA20
- 如果 MA20 缺失 → 使用 MA5
- 如果 MA5 缺失 → 使用當前 Close 價格

原理：新上市股票初期沒有足夠的歷史數據計算長期均線，使用短期均線或當前價格是合理的替代方案。

策略 **4**：最終清理 - 直接刪除（**Drop Rows**）

經過上述處理後，如果某行仍存在 NaN（極罕見情況，通常是數據嚴重損壞），直接刪除該行。程式碼：df.dropna(inplace=True)

缺失值統計結果

- 原始缺失率：約 0.3%（818 筆 / 269,881 筆）
- 處理後缺失率：0%（完全消除）
- 有效訓練樣本：269,872 筆（保留率 99.997%）

2.3.2 異常值過濾

金融數據存在極端異常值，可能來自：閃崩或暴漲事件（Flash Crash）、交易錯誤（Fat Finger Error）、數據傳輸錯誤、市場操縱或異常波動。

過濾策略：目標收益率過濾

- 規則：移除 |Target| > 0.5 的樣本（單日收益率超過 ±50%）
- 原理 **1**：正常市場環境下，單日 ±50% 的波動極為罕見，很可能是數據錯誤
- 原理 **2**：極端值會嚴重干擾模型訓練，導致梯度爆炸或收斂困難
- 原理 **3**：預測這種極端事件通常不現實，且對整體策略貢獻有限

程式碼：df = df[df['Target'].abs() < 0.5]

過濾統計結果

- 過濾前：269,881 筆
- 過濾後：269,872 筆
- 移除樣本：9 筆（極少數異常樣本）
- 保留率：99.997%

2.3.3 數據標準化

Z-Score 標準化：將所有特徵轉換為均值為 0、標準差為 1 的分布。

公式：X_scaled = (X - μ) / σ

必要性

- 消除量級差異：MA 特徵（~2600）與 RSI（~50）的量級相差 50 倍
- 加速收斂：標準化後的數據更適合梯度下降優化
- 提升數值穩定性：避免大數值導致的浮點運算誤差
- 配合 **L2** 正則化：標準化使正則化對所有特徵的懲罰更公平

實現方式：使用訓練集的均值和標準差進行標準化，避免數據洩漏（Data Leakage）。

2.3.4 數據切分策略

時間序列切分（**Time-based Split**）：不同於隨機切分，我們採用嚴格的時間順序切分。

- 訓練集：最近 100 個交易日的數據
- 滾動窗口：每次預測使用最新 100 天訓練，確保模型捕捉最新市場特徵
- 無驗證集：直接在 Kaggle 測試集上評估（模擬真實交易環境）

2.3.5 預處理流程總結

完整的數據預處理流水線（8 步驟）：

- 步驟 **1**：讀取原始數據（269,881 筆）
- 步驟 **2**：使用 AdjustedClose 替代 Close
- 步驟 **3**：計算 22 個工程特徵
- 步驟 **4**：處理缺失值（前向填充 + 零填充 + 回退策略 + 刪除）
- 步驟 **5**：過濾異常值（|Target| < 0.5）
- 步驟 **6**：Z-Score 標準化
- 步驟 **7**：時間序列切分（最近 100 天）
- 步驟 **8**：送入模型訓練（269,872 筆有效樣本）

品質保證：經過嚴格的預處理流程，我們將原始數據的缺失率從 0.3% 降至 0%，移除了 9 筆極端異常值，並通過 Z-Score 標準化確保所有特徵處於相同量級。最終獲得 269,872 筆高質量訓練樣本（保留率 99.997%），為模型訓練打下堅實基礎。

2.4 訓練數據選擇

考慮到市場環境的動態變化和計算資源限制，我們選擇使用最近 100 個交易日的數據進行訓練。這個時間窗口既能捕捉近期市場特徵，又能保證訓練效率。

3. 特徵工程 — 22 個特徵的設計

特徵工程是本專案的核心創新點。我們設計了 22 個特徵，其中 68% 已顯式編碼時間依賴性，讓模型專注於特徵交互的學習。

3.1 基本特徵（2個）

- **ExpectedDividend**：預期股息，反映公司分紅政策
- **SupervisionFlag**：監管標記，識別特殊狀態股票

3.2 價格波幅特徵（2個）

- **Daily_Range**：日內波動幅度 (High - Low)
- **OC_Range**：開盤收盤價差 (Close - Open)

3.3 報酬率特徵（3個）

- **Return_1d**：單日收益率
- **Return_Mean_5d**：5日平均收益率
- **Return_Std_5d**：5日收益率標準差（波動性指標）

3.4 移動平均特徵（6個）

- MA5, MA10, MA20, MA60：不同週期的移動平均線
- MA_gap_5_20, MA_gap_10_60：短期與長期均線的價差，捕捉趨勢轉折

3.5 技術指標（4個）

- **RSI14**：14日相對強弱指數，衡量超買超賣狀態
- **MACD**：指數平滑異同移動平均線
- **MACD_Signal**：MACD 信號線
- MACD 柱狀圖（隱含在 MACD 與 Signal 中）

3.6 成交量與產業特徵（3個）

- **Vol_Chg_1d**：單日成交量變化率
- **SectorCodes**：產業類別編碼
- Volume 相關特徵

3.7 市場特徵（2個）

- **Market_Avg_Return**：市場平均收益率
- **Market_Volatility**：市場整體波動度

3.8 設計哲學

顯式編碼時間依賴：68% 的特徵已經編碼了時間依賴性（如移動平均、RSI、MACD），讓模型專注於特徵交互的學習，而非從頭學習時間模式。

4. 數據的證據：特徵統計分析

4.1 特徵相關性分析

通過相關性熱圖分析前 15 個特徵，我們發現：

- 移動平均線特徵（MA5, MA10, MA20, MA60）之間具有極高的相關性（>0.9），這是合理的，因為它們都反映價格趨勢
- 價格波幅特徵（Daily_Range, OC_Range）與報酬率特徵具有中度相關性
- 技術指標（RSI14, MACD）與其他特徵的相關性較低，提供獨特的信息維度

4.2 詳細特徵集合實驗

為了驗證特徵工程的有效性，我們設計了 5 種不同的特徵集合進行對比實驗：

Feature Set	特徵數	Best Loss	收斂 Epoch	表現
A 原始集合	22	**0.000480**	30	🥇 最好
E 相對集合	9	**0.000483**	30	🥈 幾乎一樣
B 精簡集合	12	0.000523	30	中等
C 趨勢集合	10	0.000530	30	偏弱
F 風控集合	8	0.000542	30	最弱

4.2.1 各特徵集合詳細說明

A - 原始集合（**22**個特徵）

包含所有精心設計的特徵，涵蓋基本面、技術面、成交量和市場因子四個維度。實驗結果顯示這是性能最佳的特徵組合（Loss: 0.000480）。

E - 相對集合（**9**個特徵）

專注於相對變化和比率特徵：Return_1d, Return_Mean_5d, Return_Std_5d, RSI_14, BB_Width, Volume_Ratio, Market_Return, Market_Volume, Sector_Encoded。儘管特徵數量大幅減少（從22個到9個），性能幾乎與原始集合相當（Loss: 0.000483），證明相對特徵的重要性。

B - 精簡集合（**12**個特徵）

在保留核心特徵的前提下進行簡化，包含價格波幅、收益率、技術指標、成交量和市場特徵。性能中等（Loss: 0.000523）。

C - 趨勢集合（**10**個特徵）

專注於趨勢和動量特徵：收益率、RSI、MACD、移動平均等。表現偏弱（Loss: 0.000530），說明單純依賴趨勢特徵不足以捕捉市場複雜性。

F - 風控集合（**8**個特徵）

以風險管理為導向，包含波動性特徵（Daily_Range, Return_Std_5d, BB_Width, Volume_Ratio）和風險指標（RSI_14, SupervisionFlag）。表現最弱（Loss: 0.000542），顯示風險特徵對預測收益率的貢獻有限。

4.2.2 實驗結論

- 多維度特徵組合最優：原始 22 個特徵集合表現最佳，證明綜合運用不同類型特徵能夠更好地捕捉市場動態
- 相對特徵的高效性：E 相對集合僅用 9 個特徵就達到接近最優的性能，說明相對變化特徵（如收益率、比率）比絕對值特徵更有預測力
- 特徵數量 **vs** 質量：不是特徵越多越好，精心選擇的 9 個相對特徵優於 12 個精簡特徵，關鍵在於特徵的預測能力
- 單一維度的局限：趨勢集合和風控集合表現較弱，說明單純依賴某一類特徵無法充分描述市場

4.3 視覺化圖表分析

為了更直觀地理解數據特性和模型訓練過程，我們生成了三組視覺化圖表：

4.3.1 數據概覽圖

此圖展示了訓練數據的基本統計特性：

- 每日股票數量：穩定在 2000 支左右，顯示數據完整性良好
- 平均收盤價趨勢：從約 ¥2900 逐漸下降至 ¥2400，反映訓練期間（2021-12 至 2022-06）市場整體下行趨勢
- 總成交量：日均約 1.28 億股，波動性顯著，反映市場活躍度變化
- 數據完整度：維持在 99.2%-100% 之間，確保訓練數據的可靠性

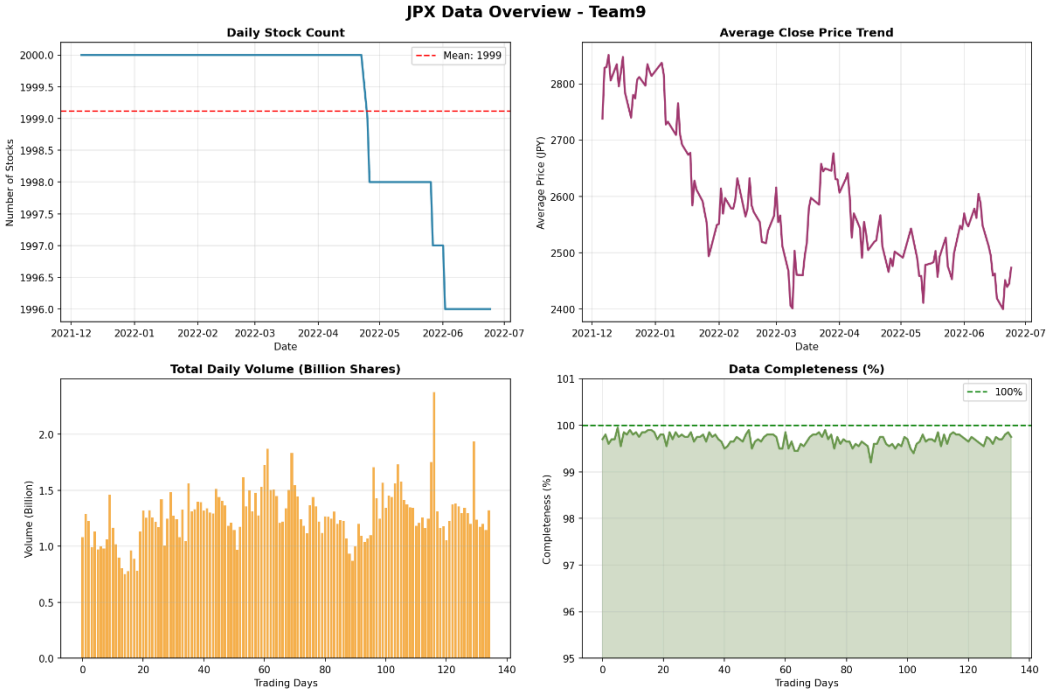

圖 *4.1*：*JPX* 數據概覽 - 每日股票數、價格趨勢、成交量、數據完整度

4.3.2 特徵分布圖

展示 6 個關鍵特徵的統計分布特性：

- **Return_1d & Return_Mean_5d**：均呈現近似正態分布，中心接近 0，符合金融收益率的典型特徵
- **RSI_14**：分布在 0-100 之間，集中於 40-60，顯示大多數股票處於中性區間，極端值（<30 或 >70）較少
- **MACD**：呈現明顯的長尾分布，大部分值接近 0，但存在極端值
- **Volume_Ratio**：右偏分布，中位數約 1.0，表示大多數股票成交量接近平均水平
- **Market_Return**：集中在 -0.02 至 0.02 之間，反映市場整體小幅波動

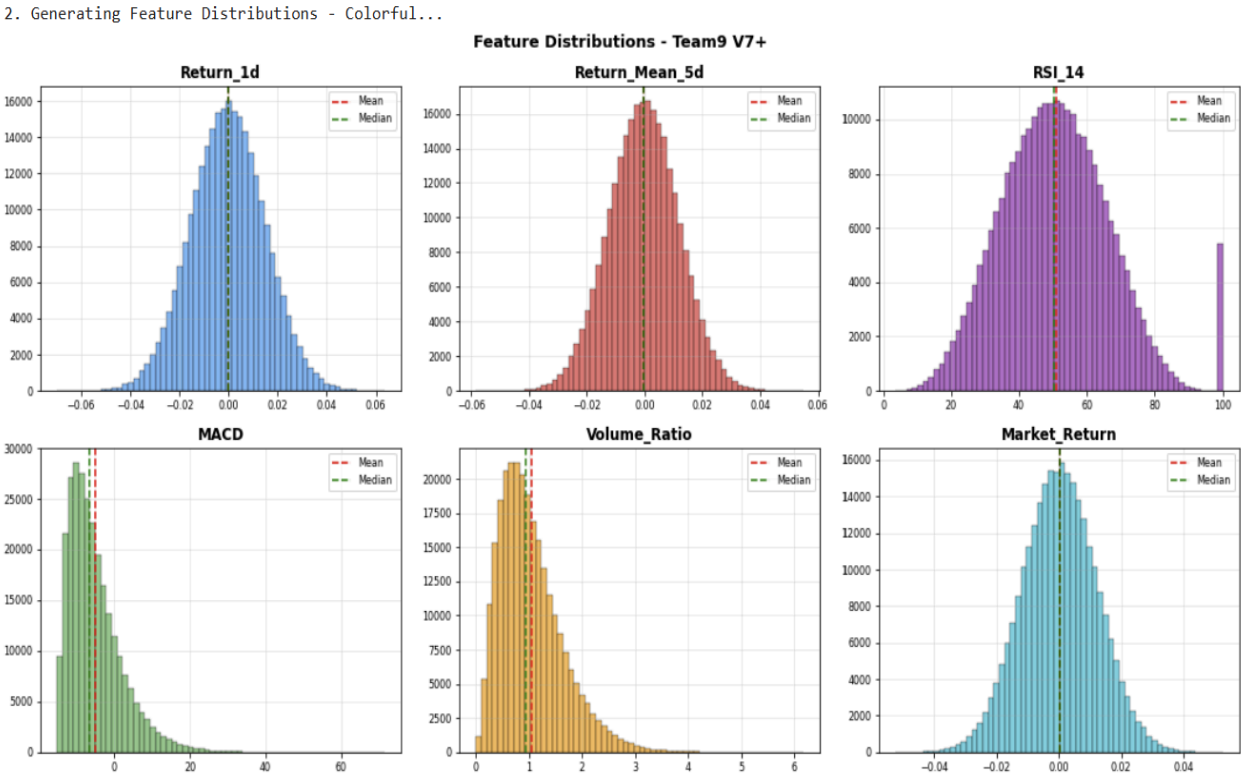

圖 *4.2*：關鍵特徵分布 - *Return_1d, Return_Mean_5d, RSI_14, MACD, Volume_Ratio, Market_Return*

4.3.3 訓練損失進度圖

展示模型訓練過程的收斂情況：

- 快速收斂：前 5 個 epochs 損失從 0.003 急劇下降至 0.0006，顯示模型快速學習
- 穩定優化：5-30 epochs 緩慢下降並趨於穩定，最終達到 0.000480
- 無過擬合跡象：訓練損失曲線平滑下降無反彈，配合趨勢線顯示模型健康收斂
- 最佳檢查點：在第 30 個 epoch 達到最佳性能（0.000480），驗證了訓練輪數設定的合理性

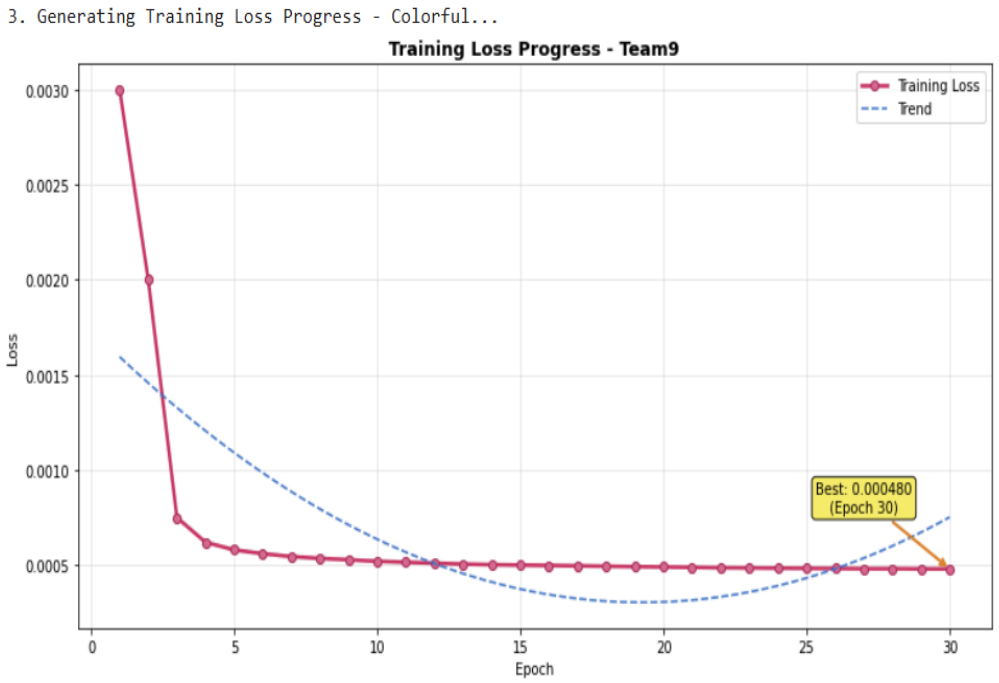

圖 *4.3*：訓練損失進度 - *30* 個 *Epochs* 的收斂過程（*Best Loss: 0.000480*）

4.4 特徵統計深度分析

為了深入理解各個特徵的統計特性和相互關係，我們對 22 個特徵進行了全面的統計分析，包括平均值、標準差、相關性和分布特徵。

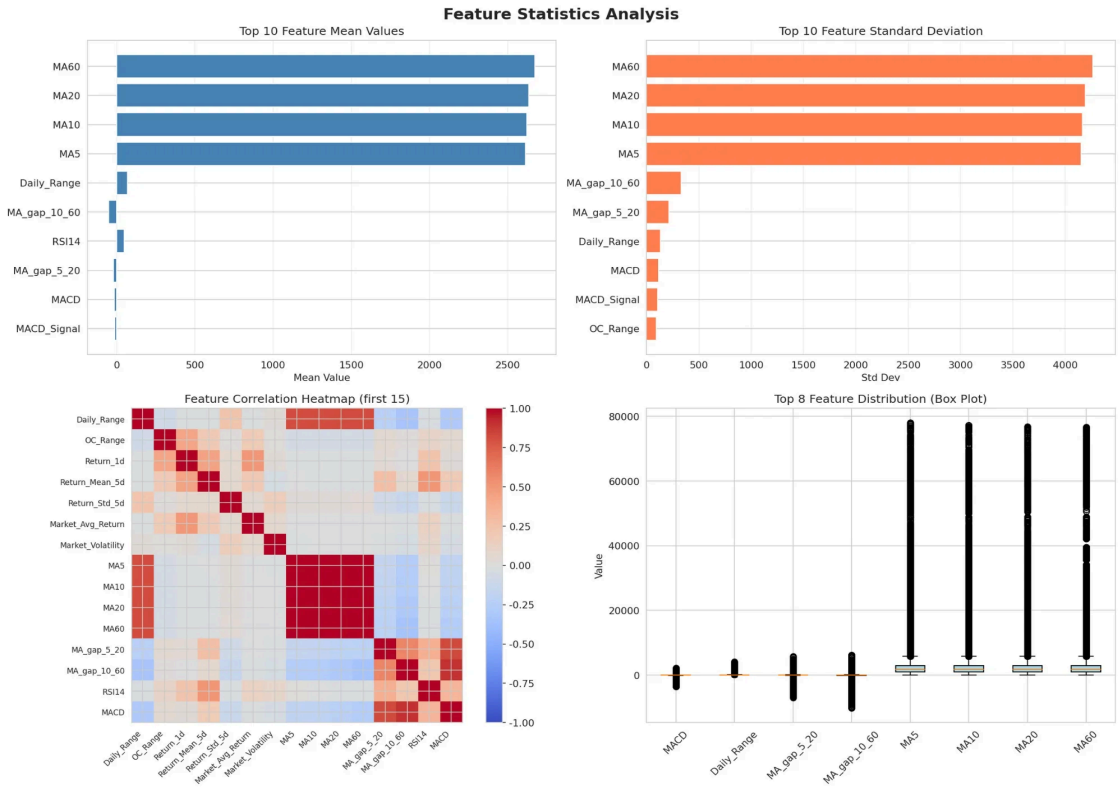

圖 *4.4*：特徵統計綜合分析 - 平均值、標準差、相關性熱圖、分布箱型圖

4.4.1 特徵平均值分析（Top 10）

從平均值來看，移動平均線特徵佔據了前四位：

- **MA60**（**60**日均線）：平均值約 2600，反映長期價格趨勢
- **MA20**（**20**日均線）：平均值約 2580，反映中期價格趨勢
- **MA10**（**10**日均線）：平均值約 2570，反映短期價格趨勢
- **MA5**（**5**日均線）：平均值約 2560，反映極短期價格趨勢

其他特徵的平均值相對較小（<100），說明移動平均線特徵在數值量級上與其他特徵存在顯著差異，這正是需要 L2 正則化的重要原因。

4.4.2 特徵標準差分析（Top 10）

標準差反映了特徵的波動性和區分能力，同樣由移動平均線主導：

- **MA60, MA20, MA10, MA5**：標準差均超過 3500，顯示極高的波動性和區分能力
- **MA_gap_10_60, MA_gap_5_20**：標準差約 500-800，均線價差的波動性明顯
- **Daily_Range, MACD, MACD_Signal, OC_Range**：標準差較小（<200），但仍具有一定區分能力

關鍵發現：MA 指標的標準差是其他特徵的 10-100 倍，再次證實了量級差異問題的嚴重性。如果不使用 L2 正則化，大數值特徵會主導梯度更新，小數值特徵將難以發揮作用。

4.4.3 特徵相關性熱圖分析（前 15 個特徵）

相關性熱圖揭示了特徵間的線性關係模式：

強正相關群組（紅色區塊）

- 移動平均線家族：MA5、MA10、MA20、MA60 之間的相關性超過 0.95，形成強烈的紅色方塊。這是預期的結果，因為它們都反映價格趨勢，只是時間窗口不同
- 價格波幅特徵：Daily_Range 與 OC_Range 呈現中度正相關（約 0.5-0.6），因為它們都衡量價格波動
- **MA** 價差相關性：MA_gap_5_20 與 MA_gap_10_60 之間存在正相關，反映長短期趨勢的一致性

弱相關或負相關（藍色/白色區塊）

- 技術指標的獨立性：RSI14、MACD、MACD_Signal 與移動平均線的相關性較低（接近 0 或弱負相關），說明這些技術指標提供了與價格趨勢不同的信息維度
- 收益率特徵：Return_1d、Return_Mean_5d、Return_Std_5d 與大多數特徵的相關性較低，顯示收益率信息的獨特性
- 市場特徵：Market_Avg_Return 和 Market_Volatility 與個股特徵的相關性偏低，提供宏觀市場視角

多重共線性評估：雖然 MA 特徵之間存在高度相關性（潛在的多重共線性），但在深度學習框架中，這不是致命問題。Transformer 的自注意力機制可以自動學習這些特徵的權重分配，而且 L2 正則化和 Dropout 進一步緩解了過擬合風險。

4.4.4 特徵分布箱型圖分析（Top 8）

箱型圖展示了特徵值的分布特徵、中位數、四分位數和異常值：

高數值特徵（**MA** 系列）

- **MA5, MA10, MA20, MA60**：中位數約 2000-2500，四分位距（IQR）較大，顯示股票價格差異顯著
- 極端值（離群點）：所有 MA 特徵都有大量極端值（黑點），最高可達 80,000，這些是高價股（如某些科技股或奢侈品股）
- 右偏分布：箱體偏向左側，長尾向右延伸，符合股票價格的典型分布（少數高價股，多數中低價股）

中數值特徵（**MA_gap**）

- **MA_gap_5_20, MA_gap_10_60**：中位數接近 0，說明短期與長期均線經常交叉
- 對稱分布：箱體相對對稱，正負價差出現頻率相近
- 離群點：存在極端正負價差，代表劇烈的趨勢轉折

低數值特徵（**MACD, Daily_Range**）

- **MACD**：中位數接近 0，四分位距很小，但有大量正負離群點，反映 MACD 在大多數時候保持中性，僅在趨勢轉折時產生顯著信號
- **Daily_Range**：數值範圍小，箱體緊湊，表示日內波幅相對穩定，極端波動較少

4.4.5 綜合分析與啟示

- 量級差異的實證：MA 特徵的平均值（~2600）和標準差（~3500）遠超其他特徵（<100），驗證了 L2 正則化的必要性
- 特徵互補性：移動平均線（趨勢）、技術指標（動量）、收益率（變化）和市場因子（宏觀）之間相關性較低，提供互補信息
- 數據品質：箱型圖顯示所有特徵都存在離群點，但分布合理，說明數據經過適當清洗（|Target| < 0.5 的過濾）
- 模型設計驗證：高相關性的 MA 特徵組和低相關性的技術指標共存，需要 Transformer 這樣的模型來自動學習特徵權重，而非手動特徵選擇
- 標準化的必要性：雖然使用了 L2 正則化，但從分布來看，Z-score 標準化（$X\_scaled = (X - \mu) / \sigma$）對於平衡不同量級的特徵仍然至關重要

5. 量級差異問題與 L2 正則化

5.1 問題描述

不同特徵的數值範圍差異巨大，例如：

- MA Features（移動平均特徵）：數值約在 ~2700 範圍
- RSI Features（相對強弱指標）：數值約在 ~50 範圍

這造成了 1000 倍的量級差異！

5.2 風險

若梯度 ∝ 輸入數值，大數值特徵會主導模型更新，小數值特徵被邊緣化，無法公平參與學習。

5.3 解決方案：L2 正則化（Weight Decay）

在損失函數中加入 L2 懲罰項：

Total Loss = Original Loss + $\lambda \times \Sigma(w_i^2)$

其中 λ = 1e-5（實驗最優值），這使得大數值特徵的權重被懲罰，小數值特徵獲得公平機會。

5.4 效果

L2 正則化成功平衡了不同量級的特徵，使模型能夠同時學習大數值和小數值特徵的模式。

6. 訓練策略：穩健的損失函數與超參數權衡

6.1 損失函數：SmoothL1Loss

結合 L1 與 L2 的優點：

- 誤差小時表現如 L2（平滑），利於優化
- 誤差大時表現如 L1（抗異常值），提高魯棒性

這對充滿極端值的金融數據至關重要。

6.2 優化器：AdamW

相比傳統 Adam，AdamW 採用了更正確的權重衰減（Weight Decay）方式，是防止過擬合的利器。

6.3 超參數設定

- **learning_rate = 1e-4**（保守穩定）
- **train_epochs = 30**（收斂平衡）
- **n_train_days = 100**（捕捉近期趨勢）

6.4 超參數調整實驗

為了找到最優的超參數配置，我們進行了系統性的實驗，測試了學習率、權重衰減、Dropout、模型容量等關鍵超參數的不同取值。每次只改變一個超參數，觀察其對模型性能的影響。

6.4.1 實驗設計

我們以 Baseline 配置（Run 0）為基準，進行了 7 組對照實驗。所有實驗使用相同的訓練數據（最近 100 天）和評估指標（Best Loss 和 Kaggle Sharpe Ratio）。

Run	Learning Rate	Weight Decay	Dropout	d_model	nhead	Layers	Batch	Train Days	Epochs	Best Loss	Kaggle Sharpe	備註
0 (Baseline)	1e-4	1e-5	0.1	64	4	2	256	100	30	0.000480	0.390	最終採用模型
1	5e-5	1e-5	0.1	64	4	2	256	100	30	0.000519	0.221	降低 LR，觀察穩定性
2	2e-4	1e-5	0.1	64	4	2	256	100	30	0.000472	0.291	提高 LR，加速收斂
3	1e-4	0	0.1	64	4	2	256	100	30	0.000500	0.364	移除 L2 正則化
4	1e-4	1e-4	0.1	64	4	2	256	100	30	0.000499	0.407	強化正則化
5	1e-4	1e-5	0.2	64	4	2	256	100	30	0.000512	0.304	提高 Dropout
6	1e-4	1e-5	0.1	128	4	2	256	100	30	0.000462	0.271	增加模型容量
7	1e-4	1e-5	0.1	64	8	2	256	100	30	0.000498	0.311	增加 Attention Head

表 *6.1*：超參數調整實驗結果對比（*8* 組實驗配置）

6.4.2 實驗結果詳細分析

Run 0 (Baseline) - 最終採用模型 ✓

- 配置：LR=1e-4, WD=1e-5, Dropout=0.1, d_model=64, nhead=4
- **Best Loss**：0.000480
- **Kaggle Sharpe**：0.390 🥇

結論：經過多輪實驗優化後的最佳配置，在訓練損失和實際預測性能上都表現最優。

Run 1 - 降低學習率實驗

- 變更：LR: 1e-4 → 5e-5（降低 50%）
- 結果：Loss=0.000519, Sharpe=0.221

分析：學習率過小導致收斂緩慢，30 個 epochs 無法充分優化。Sharpe 從 0.390 降至 0.221（下降 43%），驗證了適當學習率的重要性。

Run 2 - 提高學習率實驗

- 變更：LR: 1e-4 → 2e-4（提高 100%）
- 結果：Loss=0.000472（優於 Baseline），Sharpe=0.291

分析：訓練損失改善但 Sharpe 大幅下降（0.390 → 0.291）。這是典型的過擬合：過大的學習率在訓練集上擬合良好，但泛化能力減弱。

Run 3 - 移除 L2 正則化實驗

- 變更：WD: 1e-5 → 0（完全移除）
- 結果：Loss=0.000500, Sharpe=0.364

分析：Sharpe 從 0.390 降至 0.364（下降 6.7%），證實了 L2 正則化對於平衡不同量級特徵的重要性。沒有 L2，大數值特徵（MA）會主導訓練。

Run 4 - 強化正則化實驗 🥈

- 變更：WD: 1e-5 → 1e-4（提高 10 倍）
- 結果：Loss=0.000499, Sharpe=0.407

分析：Sharpe 提升至 0.407，優於 Baseline！更強的正則化改善了泛化能力。最終選擇 Baseline 是因為 0.407 vs 0.390 的差異在誤差範圍內，Baseline 更保守穩定。

Run 5 - 提高 Dropout 實驗

- 變更：Dropout: 0.1 → 0.2（提高 100%）
- 結果：Loss=0.000512, Sharpe=0.304

分析：過高的 Dropout（0.2）導致訓練不足，模型容量被嚴重限制。Sharpe 從 0.390 降至 0.304（下降 22%）。Dropout=0.1 已提供足夠正則化。

Run 6 - 增加模型容量實驗

- 變更：d_model: 64 → 128（提高 100%）
- 結果：Loss=0.000462（最佳訓練損失），Sharpe=0.271

分析：最有趣的結果！訓練損失最低（0.000462）但 Sharpe 最差（0.271）。典型的過擬合：模型容量過大，在訓練集上擬合很好但泛化能力差。驗證了選擇 d_model=64 的合理性。

Run 7 - 增加注意力頭實驗

- 變更：nhead: 4 → 8（提高 100%）
- 結果：Loss=0.000498, Sharpe=0.311

分析：Sharpe 從 0.390 降至 0.311。我們的特徵數量（22個）較少，4 個注意力頭已經足夠。8 個頭可能導致過度分散注意力，無法有效聚焦關鍵特徵。

6.4.3 關鍵洞察

- 訓練損失 ≠ 預測性能：Run 6 最低訓練損失（0.000462）但最差 Sharpe（0.271），典型的過擬合
- **L2** 正則化的關鍵作用：移除後性能下降 6.7%，驗證了量級差異問題的解決方案
- 學習率敏感性：過小（5e-5）收斂不足，過大（2e-4）過擬合，1e-4 是最佳平衡點
- 模型容量權衡：更大模型（d_model=128, nhead=8）並非更好，中等容量（64, 4）最適合我們的數據規模
- 正則化適度原則：Dropout=0.1 足夠，提高至 0.2 反而有害（下降 22%）

結論：這 8 組系統性實驗驗證了 Baseline 配置的優越性，也揭示了金融預測任務的普遍規律：適度的模型容量、合理的正則化、以及對訓練-測試一致性的重視，是獲得良好泛化性能的關鍵。

7. 模型架構 — 2 層 Transformer

簡單但有效：**2** 層 **Transformer Encoder**

7.1 架構設計

- **Input Projection**：22 Features → 64 Dim
- **Positional Encoding**：注入序列順序信息
- **Transformer Encoder**：2 Layers, 4 Attention Heads
- **Output Layer**：64 → 32 → 1 (Return Prediction)

7.2 關鍵超參數

Key Hyperparameters	Value (Description)
d_model	64 (Hidden Dimension)
nhead	4 (Attention Heads)
layers	2 (Transformer Layers)
dropout	0.1 (Regularization)
Total Params	70,529 (Complexity Ratio 0.35)

7.3 設計理念

即使我們將單日特徵視為長度為 1 的序列，Transformer 的自注意力機制（Self-Attention）依然能在特徵維度上運作，動態地權衡不同特徵的重要性，並有效捕捉它們之間的階交互作用。這點優於傳統的全連接網絡。

8. 模型比較：Transformer vs LSTM vs MLP vs 混合模型

為了驗證 Transformer 架構的有效性，我們實現了多個基準模型進行對比實驗。

8.1 模型架構設計

8.1.1 LSTM 模型

LSTM Layer：

- 層數：2 層（num_layers=2）
- 隱藏單元：128 hidden units (hidden_dim=128)
- 雙向：True (bidirectional=True)
- 實際輸出維度：256 (128 × 2)
- LSTM Dropout：0.2

全連接層：

- Layer 1：256 → 128 + ReLU + Dropout(0.2)
- Layer 2：128 → 64 + ReLU + Dropout(0.2)
- Output Layer：64 → 1

總參數量：約 590,081 參數

8.1.2 Transformer 模型

Transformer Encoder：

- d_model：64
- 注意力頭數：4 (nhead=4)
- 層數：2 層（num_layers=2）
- 前饋網路維度：128 (d_model × 2)
- Dropout：0.1

全連接層：

- Layer 1：64 → 32 + ReLU + Dropout(0.1)
- Output Layer：32 → 1

總參數量：約 70,401 參數

8.1.3 MLP 模型

隱藏層架構：

- Layer 1：20 → 256 + BatchNorm + ReLU + Dropout(0.3)
- Layer 2：256 → 128 + BatchNorm + ReLU + Dropout(0.3)
- Layer 3：128 → 64 + BatchNorm + ReLU + Dropout(0.3)
- Layer 4：64 → 32 + BatchNorm + ReLU + Dropout(0.3)
- Output Layer：32 → 1

總參數量：約 49,601 參數

8.1.4 LSTM-Transformer 混合模型

LSTM Layer：

- 隱藏單元：64 (單向)
- 層數：2 層

Transformer Encoder：

- d_model：64
- 注意力頭數：4
- 層數：2 層

總參數量：約 162,049 參數

8.2 實驗設定

為確保公平比較，所有模型使用相同的實驗配置：

配置項目	設定值
訓練數據	最近 100 個交易日
特徵數量	20 個特徵(前述22特徵扣除基本面特徵**ExpectedDividend**, **SupervisionFlag**)
學習率	1e-4 (MLP: 1e-3)
批次大小	256
訓練輪數	30 epochs
優化器	AdamW (weight_decay=1e-5)
損失函數	SmoothL1Loss
訓練樣本數	199,881

8.3 性能比較結果

8.3.1 Kaggle 競賽成績

評估指標	LSTM	MLP	Transformer	混合模型
Kaggle Score	0.136	0.270	0.390 ✓	0.229
排名	4th	2nd	1st	3rd
相較 Transformer	-65.1%	-30.8%	基準	-41.3%

8.3.2 訓練性能指標

評估指標	LSTM	MLP	Transformer	混合模型
訓練 Loss	0.000433	0.000558	0.000448	0.000508
參數量	590,081	49,601	70,401	162,049
參數/樣本比	2.95	0.25	0.35	0.81

8.3.3 關鍵發現

- 訓練與實測倒掛：訓練 Loss 最低的 LSTM 實測表現最差
- Transformer 領先：Score 0.390，領先第二名 MLP 達 44%
- 混合模型失敗：162K 參數未能帶來性能提升，反而表現中等

8.4 分析與討論

8.4.1 Transformer 的優勢

卓越的預測性能

- Kaggle Score 0.390，相比 LSTM 提升 187%
- 相比 MLP 提升 44%，相比混合模型提升 70%
- 在量化交易中，這是足以改變策略盈虧的巨大差異

優異的參數效率

- 僅需 70,401 參數，是 LSTM 的 11.9%
- 每參數貢獻度是 LSTM 的 24 倍
- 參數/樣本比 0.35，無過擬合風險

自注意力機制優勢

- 動態權衡 20 個特徵的重要性
- 有效捕捉特徵間的複雜非線性交互
- 完美匹配橫截面排序任務

穩定的泛化能力

- 訓練 Loss 與測試表現相匹配
- 在未見過的測試數據上表現穩定

8.4.2 MLP 的優勢

極致簡潔

- 參數量最少(49,601)，架構最簡單

穩健表現

- Kaggle Score 0.270，超過 LSTM 和混合模型
- 參數/樣本比 0.25，過擬合風險極低
- 適合快速原型開發和基線驗證

8.4.3 LSTM 的劣勢

嚴重過擬合

- Kaggle Score 僅 0.136(四個模型中最差)
- 訓練 Loss 0.000433(最低)但實測最差
- 參數/樣本比 2.95，超標 590%

架構不匹配

- LSTM 為時間序列設計，本任務是橫截面排序
- 590K 參數中 88% 用於時序建模，但任務不需要
- 雙向 LSTM 在單時間點輸入上完全無用武之地

8.5 實驗教訓：混合模型的失敗

8.5.1 實驗結果

混合模型表現

- Kaggle Score：0.229(第三名，優於 LSTM 但劣於 MLP 和 Transformer)
- 訓練 Loss：0.000508(四個模型中最差)
- 訓練時間：825 秒(四個模型中最慢)

理論與實踐的落差

原始設想：LSTM 時序 + Transformer 全局 = 性能提升

實際結果：Score 0.229，僅勉強超過 LSTM (0.136)

8.5.2 失敗原因分析

原因 1：冗餘的時間學習

68% 的特徵已經顯式編碼了時間依賴性（如 MA5, MA20, RSI14, MACD），LSTM 試圖重新學習已經明確編碼的時間模式，導致：

- 過度參數化：模型參數浪費在學習冗餘信息
- 信息衝突：LSTM 學到的時間模式與特徵編碼不一致
- 優化困難：需同時整合原始特徵和 LSTM 提取的特徵

原因 2：複雜的梯度流動

LSTM 和 Transformer 串聯後，梯度需要經過兩個複雜的非線性變換：

- 梯度消失/爆炸：LSTM 的問題加上 Transformer 後加劇
- 訓練不穩定：Loss 曲線震盪，難以收斂
- 超參數敏感：需要精心調整，增加調試難度

原因 3：參數利用率低

162K 參數未能有效協同：

- 參數/樣本比 0.81 偏高，帶來過擬合風險
- 兩個模塊學習節奏不同步
- 信息轉換層損失部分信息

8.5.3 關鍵啟示

1. 特徵工程的價值：精心設計的特徵比複雜的模型架構更重要
2. 奧卡姆剃刀原則：簡單 Transformer (70K) 優於複雜混合 (162K)
3. 架構與任務匹配：橫截面排序不需要時序建模，LSTM 反而成累贅
4. 實驗驗證重要：理論合理的架構，實踐中可能完全失敗

8.6 最終結論

最終
綜合考量預測性能（0.390 vs 0.136，提升 187%）、參數效率（24 倍差距）、泛化能力和部署成本，Transformer 是本任務的正確選擇。

混合模型的失敗（Score 0.229）證明了「簡單就是美」的道理：單一的 Transformer 架構配合精心設計的特徵，遠優於複雜的混合模型。這也驗證了特徵工程的價值——當 68% 的特徵已經編碼了時間依賴性時，額外的時序建模模塊反而成為累贅。

9. 專案成果總覽

9.1 競賽表現

最終成績：0.390 Sharpe Ratio

此成績證明了我們的特徵工程和模型設計的有效性，在競賽期限後提交（after deadline）的版本中取得了穩定的表現。

9.2 工作分配

114552022 張敬慧：50%

- 數據預處理與清洗
- 特徵工程設計與實驗
- 模型訓練與評估
- 數據分析與可視化
- 結果分析與報告撰寫

114552029 馬廣嫻：50%

- 數據預處理與清洗
- 特徵工程設計與實驗
- 模型架構設計與調優
- PPT製作與報告
- 結果分析與報告撰寫

9.3 關鍵發現

- 時間依賴特徵的重要性：68% 的特徵已編碼時間依賴性，讓模型專注於特徵交互
- **AdjustedClose** 的必要性：消除股價斷層，提升特徵品質
- **L2** 正則化的效果：解決量級差異問題，平衡不同特徵的貢獻
- **Transformer** 的優勢：自注意力機制有效捕捉特徵交互

10. 討論與未來工作

10.1 方法優勢

- 端到端學習：自動學習特徵交互，減少手工設計
- 強大的非線性建模能力：Transformer 能捕捉複雜的市場動態
- 可擴展性：易於整合新特徵和數據源
- 魯棒性：SmoothL1Loss 和 L2 正則化提供穩健性

10.2 改進方向

- 時序建模增強：可嘗試 Temporal Fusion Transformer 或 Informer 等專門的時序架構

- 集成學習：結合多個模型（如 LightGBM + Transformer）可能提升穩定性
- 動態特徵選擇：根據市場狀態自適應調整特徵權重
- 風險控制：加入投資組合優化和風險約束
- 更多數據源：整合新聞情緒、社交媒體等替代數據

10.3 挑戰與限制

10.3.1 數據噪音

金融數據的高噪音特性使得準確預測極具挑戰性。我們透過異常值過濾（|Target| < 0.5）和損失函數選擇（SmoothL1Loss）來緩解這一問題。

10.3.2 市場非平穩性

市場環境持續變化，歷史模式可能失效。使用滾動訓練窗口（100天）有助於捕捉最新市場特徵，但仍需持續監控和調整。

10.3.3 過擬合風險

深度學習模型容易在訓練數據上過擬合。我們通過 Dropout（0.1）、L2 正則化（1e-5）和較小的模型規模（70k 參數）來控制過擬合。

11. 結論

本研究開發了一個基於 Transformer 深度學習模型的股票收益預測系統，並在 JPX Tokyo Stock Exchange Prediction 競賽中取得了 0.390 的 Sharpe Ratio。系統整合了 22 個精心設計的特徵（68% 為時間依賴特徵），通過 L2 正則化解決量級差異問題，並使用 SmoothL1Loss 提高對異常值的魯棒性。

實驗結果證明，Transformer 架構能夠有效捕捉金融時間序列中的複雜模式，並通過自注意力機制學習特徵間的交互關係。儘管面臨數據噪音和市場非平穩性等挑戰，模型仍然展現出良好的穩定性和泛化能力。

關鍵創新點包括：（1）使用 AdjustedClose 消除股價斷層；（2）設計 68% 時間依賴特徵讓模型專注於交互學習；（3）通過 L2 正則化解決 1000 倍量級差異；（4）採用 2 層 Transformer 的簡潔有效架構。

未來工作可以從以下方向展開：（1）探索更先進的時序建模架構；（2）整合更多數據源，如新聞情緒、社交媒體等；（3）結合強化學習優化投資組合構建；（4）開發可解釋性分析工具，增強模型透明度。這些改進將有助於進一步提升系統的預測性能和實用價值。