Kaggle 比賽報告: Store Sales - Time Series Forecasting

😉 學制班組

四技資管三由

11 組員

- 11146001 何國瑰
- 11146010 楊承憧
- 11146027 高偉倫

🖈 專案摘要

本報告為參與 Kaggle 所舉辦之「Store Sales - Time Series Forecasting」競賽的期末專題,預測厥瓜多各商店每日的銷售額。

我們使用 pandas 進行資料處理,並以 PyTorch 建立 MLP 與 LSTM 模型進行時間序列預測。模型表現以 RMSE (Root Mean Squared Error) 評估,最終透過 五折交叉驗證 獲得穩定預測。

🏲 資料集說明

使用官方提供之資料如下:

train.csv : 訓練資料,每日每店每商品類別之銷售資訊

test_csv : 需預測的測試資料集

oil.csv : 每日油價資訊 (DCOILWTICO)

holidays_events.csv : 節日與活動資訊

transactions.csv :每日店鋪交易數量

stores.csv:商店資訊(城市、州、省份、群集編號等)

資料預處理與特徵工程 (preprocess.py)

主要步驟:

- 1. 合併所有外部資訊(油價、節慶、交易量、店鋪資料)
- 2. 去除 Transfer 節慶類型,新增 is_holiday 欄位
- 3. 新增時間特徵欄位:
 - dow, month, year, weekofyear, is_weekend
- 4. 編碼類別變數:
 - family_enc, type_enc, cluster_enc
- 5. 建立目標欄位:
 - sales_log = log1p(sales)
- 6. 建立 Lag 特徵:
 - sales_lag_1, sales_lag_7, sales_lag_14, sales_lag_28

最終輸出檔案:

- train_basic.parquet
- test_basic.parquet

使用特徵:

```
"onpromotion", "dcoilwtico", "transactions", "is_holiday",
  "dow", "month", "year", "weekofyear", "is_weekend",
  "family_enc", "type_enc", "cluster_enc",
  "sales_lag_1", "sales_lag_7", "sales_lag_14", "sales_lag_28"
]
```

模型設計

夕 模型一 (MLP)

- 入層: 特徵數 → 128 → ReLU + BatchNorm
- 隱藏層: 64 → ReLU + BatchNorm
- 輸出層: 1

副練流程 (train_model.py)

- Dataset: 自定義 SalesDataset
- 分群: GroupKFold(n_splits=5) 以 store nbr 分群

- 損失函數: MSELoss
- 每 fold 獨立儲存最佳模型
- 最終以對於 test 的預測加算平均

配置

Epochs: 20

Batch Size: 1024Optimizer: Adam

Learning Rate: 0.001

₹ 進階模型二 (LSTM)

⋠ LSTM 構造

Input → LSTM(2層, hidden=64) → FC(1)

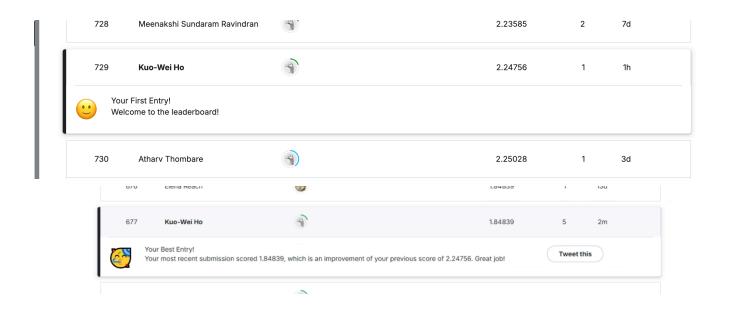
- 使用 PyTorch 建立 SeqSalesDataset
- Sliding window 列列設計 sequence length = 7
- 訓練運行方式與 MLP 相同
- 每 fold 保留 best RMSE 的模型
- 預測時先預留 7 個 padding 作為系列後部
- 預測值 expm1() 還原

↑ 結果與提交

- 測試預測值 → log_pred 還原 → sales = expm1(log_pred)
- 將預測結果儲存為 submission.csv

🃅 成果截圖

本競賽最終成績



显 技術說明

- 系統環境管理使用 uv
- 資料處理使用 pandas
- 算法層使用 PyTorch 建構 MLP 與 LSTM
- 試過不同特徵組合與平均 ensemble
- 以圖示和 RMSE 評估表現模型效能

☑ 結論

本專題由組員分工合作,自行寫程與設計預測流程,包括資料預處理、特徵工程與建立神經網路模型,成功完成店鋪每日銷售預測工作!