Financial index prediction based on machine learning

基於機器學習進行金融指數預測

組別：A89 組員：何知諺 指導教授：翁詠祿

Abstract

近年來機器學習興起，應用層面由影像處理、醫療用途到金融產業皆能見其蹤影。本專題基於機器學習模型進行金融指數預測，旨在設計一套「股價預測系統」，其功能包含讀取股票資訊、股價時間軸切割、資料預處理、股票預測及資料視覺化。

以精準預測股價為目標，試圖解釋不同機器學習模型預測結果的差異。取得股票歷史數據後透過設計好的系統切割時間序列，接著使用資料科學及統計學理論進行資料預處理，再將經過前處理的資料送進不同的機器學習模型如:Polynomial regression、XGBoost、LSTM，觀察訓練後的模型是否能順利測試資料。模型得依據過去五天的股票資訊，預測後一天股價數值，最後使用 MSE 評估預測數值和實際股價的距離。實驗結果顯示，機器學習模型確實能準確預測未來股價。期望研究成果最終能應用於交易市場，提供強度更高的股票交易策略。

Introduction

本專題嘗試使用機器學習預測股票市場走向，由財金平台 Investing.com 取得金融歷史數據，使用 polynomial regression、XGBoost regression、LSTM regression 進行迴歸分析，並透過 Python 中相關 package 完成資料預處理，期望能以過去資訊精準畫出股票未來走勢圖，將研究成果運用於交易程式。

一、 系統設計

1.1 資料愈處理

A.標準化(Standardization)

建立機器學習模型時，會使用梯度下降法(Gradient Descent)計算最佳解，若梯度等高線為窄長型將造成收斂緩慢甚至無法收斂的情形。透過 feature 標準化，將梯度等高線圖轉換成同心圓，得增加收斂

速度，大幅降低機器學習模型訓練時間。此外，各個 feature 間的數值大小範圍不同，造成訓練時無意間提高比重，透過標準化資料，此問題得以解決，更能提高模型表現。

B. Polynomial Features transform

許多機器學習模型如 support vector machine、logistic regression、Linear regression 皆為線性，無法擬合非線性資料。顯然地，股票數據並非線性，為了讓線性模型得以使用，其一做法是將 feature 投影到更高維度，使其呈線性分布，或是加入各個 feature 的高次方項作為新的 feature。

1.2 機器學習模型

A. polynomial regression

線性模型進行的預測方式，是計算輸入特徵的加權總合，再加上偏差項(bias)的常數。將原特徵項的高次項加入，作為新的特徵項，即為多項式迴歸。模型訓練過程採用梯度下降(Gradient Decent)，它可以幫助模型找到最佳參數，以線性迴歸為例，透過 training data 進行梯度下降，模型得找出參數向量的最佳數值。最後透使用 MSE 評估模型對 test data 的表現。

B. XGBoost

Boost 泛指將弱學習模型組合成強學習模型的方法，本次使用的 XBGoost(Extreme Gradient Boosting)，組合多棵迴歸樹決定 sample 最後預測結果，本演算法由 C++撰寫因此速度快。每棵迴歸樹的節點由 sample 的 feature 決定，葉子則是 sample 在這棵樹得到的分數，最後將每一棵樹的分數相加。

C. LSTM

深度學習模型—RNN(Recurrent Neural Network)，在處理時間序列資料時，會遺失先前的資訊，為了解決這個問題，有了 LSTM(Long Short Term Memory networks)長短期記憶模型的提出。在 LSTM 細胞中，input gate 和 forget gate 皆會根據當前輸入和先前狀態學習辨識重要輸入，並且將其存入長期狀態，以取得更多脈絡。本專題目標藉機器學習分析過往股票資訊與預測未來走勢，LSTM 是當前分析時間序列資料常見的深度學習模型，非常值得進行實作並予以探討。

二、 實驗結果

取得股票資訊後進行資料愈處理，基於 polynomial regression、XGBoost、LSTM 三種 model 對元大電子科技基金、中華電信、S&P500 三檔股票進行預測並計算 test data MSE。

表 1 模型預測結果

股票 / Model	元大電子科技基金	中華電信	S&P500
polynomial regression	0.01073	0.02177	0.00497
XGBoost	0.05043	0.02795	0.09414
LSTM	0.01931	0.02168	0.03878

模型對三檔股票的預測結果一致，polynomial regression 可得到最佳結果，最為適合分析股票類別的線圖。如同預期，雖然 XGboost 在分類標籤方面的應用能有絕佳的表現，但在迴歸分析中表現並非突出。LSTM 亦能有效 fit 模型，相對於線性迴歸單純的數學式，它的運算更為複雜，模型可改良的空間更多，未來的研究得持續優化。各模型表現結果可能也與資料預處理的方式有所關聯。

三、 研究結果

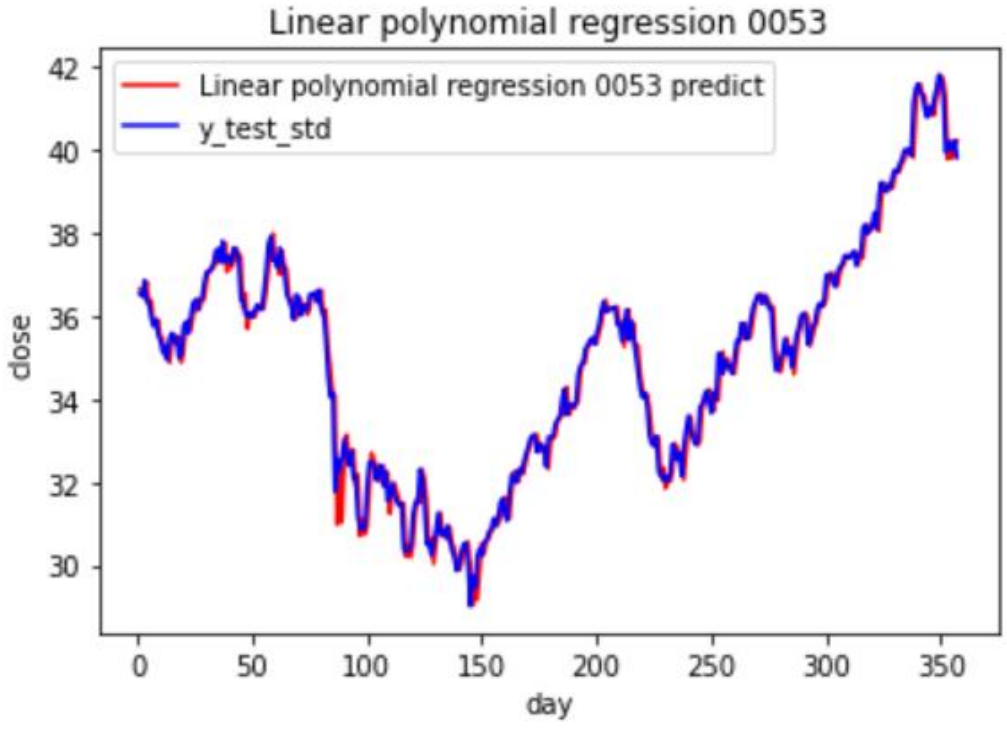

圖 1polynomial regression 預測元大電子基金股價走勢

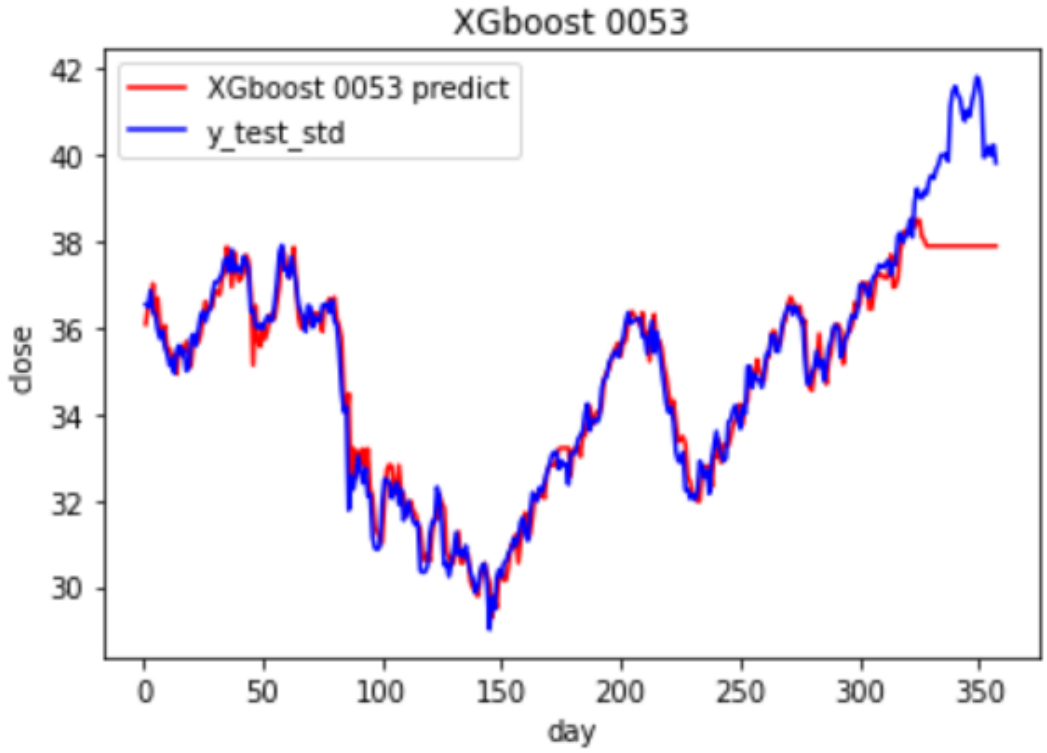

圖 2 XGBoost 預測元大電子基金股價走勢

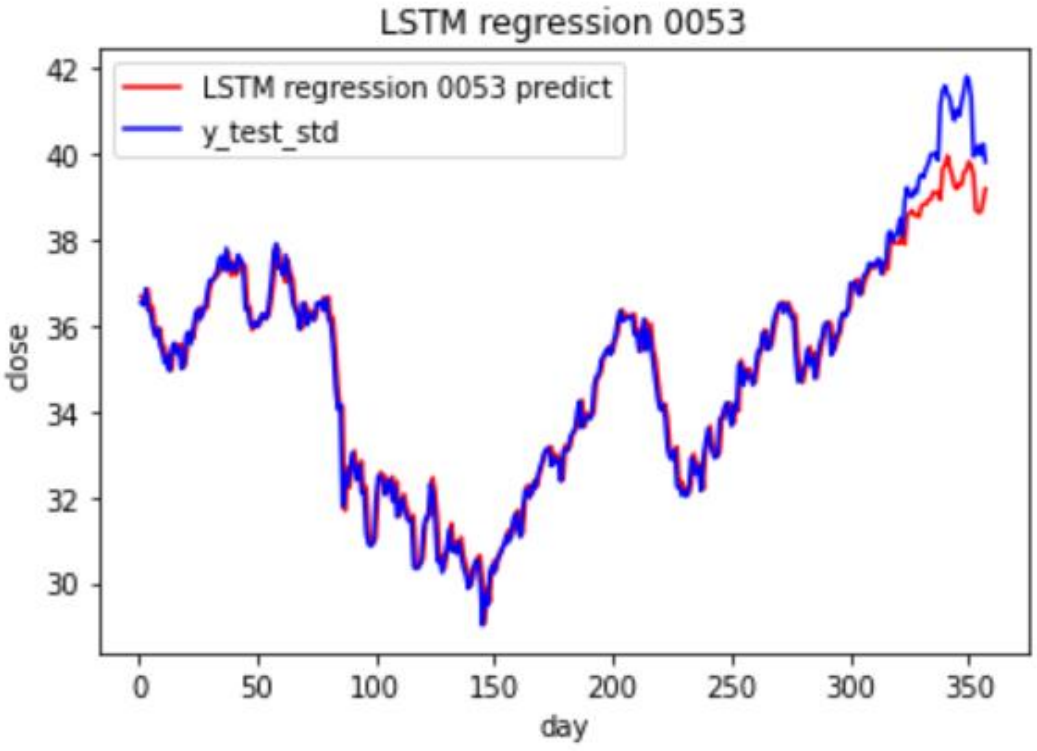

圖 3 LSTM 預測元大電子基金股價走勢

心得

為期一年的專題實作，使我習得多項新知。起初，看著國內外開放式課程自學 python、機器學習、股票知識而後進行整合，過程中時常遇上困難，卻因此提升解決問題的能力，尤其如何善用網路資源，並在龐大的資訊中理出思緒、汲取所需。也因為這些經驗，我得以淺嚐研究生的生活，體驗在課業中同時進行研究並且發現問題、解決問題。對於學長和教授撥冗給予專題指導不勝感激，期許自己未來在特定領域有所成後，亦能如此提攜後進。