國立成功大學

電機工程學系

碩士論文

利用深度學習預測股價趨勢反轉點-以標普500為例

Using Deep Learning to Predict Trend Reversal Points of Stock price -

A Case Study of the S&P 500

研究生: 林鉦平 Student: Chang-Ping Lin

指導教授: 吳謂勝 Advisor: Wei-Sheng Wu

Department of Electrical Engineering

National Cheng Kung University

Tainan, Taiwan, R.O.C

Thesis for Master of Science

July, 2024

中華民國一一三年七月

利用神經網絡預測金融商品價格趨勢反轉點

研究生: 林鉦平

指導教授: 吳謂勝

國立成功大學電機工程學系

**摘要**

　　市場未來的趨勢走勢及趨勢反轉點的預測一直是金融領域中的重要課題，但目前尚無絕對準確的預測方法。過去，許多學者利用技術指標分析市場狀況，但技術分析因其滯後性難以達到預測效果。隨著機器學習技術的發展，越來越多的研究開始探討利用機器學習和深度學習技術來預測未來的趨勢走勢及趨勢反轉點。對於趨勢的預測較為容易，但因難以即時預測反轉點的出現，無法制定有效的交易策略。對於趨勢反轉點的預測則較為困難，但若能準確預測則能制定有效的交易策略，為投資者帶來巨大的獲利。

　　本研究利用深度學習模型，結合趨勢預測及趨勢反轉點預測的優點，透過預測未來數日中每日的趨勢狀況，識別趨勢反轉點的出現。透過預測趨勢反轉點，為投資者建立交易策略的參考。首先定義趨勢反轉點為前後20日內之局部極值，並篩選相鄰同方向的反轉點，確保反轉點之間相鄰的一定是不同方向的反轉點後，定義趨勢反轉點之間的區間為趨勢，由高點至低點為下跌趨勢、由低點至高點為上漲趨勢。資料集以標普500原始股價加入美國公債殖利率及技術指標製備，並以滑動視窗方法切分模型的輸入及輸出。模型選用CNN、LSTM、Transformer等時間序列模型，以過去64天之32種特徵作為輸入，以未來16天之每日趨勢為輸出。結果…

關鍵字: 趨勢預測、趨勢反轉點、機器學習、深度學習

**Using Neural Networks to Predict Trend Reversal Points in Financial Instrument Prices**

Student: Chang-Ping Lin

Advisor: Wei-Sheng Wu

Department of Electrical Engineering, National Cheng Kung University

|  |
| --- |
| **SUMMARY**  　　The prediction of future market trends and trend reversal points has always been a crucial topic in the financial sector, but so far, no absolutely accurate prediction method has been proposed. In the past, many scholars have used technical indicators to analyze market conditions, but due to their lagging nature, technical analysis cannot achieve predictive results. With the development of machine learning technology, an increasing number of studies have begun to explore the use of machine learning and deep learning techniques to predict future trends and trend reversal points. While trend prediction is relatively easier, it is difficult to formulate effective trading strategies due to the challenge of real-time prediction of reversal points. Predicting trend reversal points is more difficult, but if accurate, it can lead to the development of effective trading strategies, bringing substantial profits to investors.  　　This study utilizes deep learning models to combine the advantages of trend prediction and trend reversal point prediction. By predicting the daily trend conditions over the next few days, it identifies the occurrence of trend reversal points. Accurate prediction of trend reversal points provides a valuable reference for investors to establish trading strategies.  Key word: Trend Prediction, Trend Reversal Points, Machine Learning, Deep Learning |

**INTRODUCTION**

Kawasaki disease is a condition that predominantly affects children under the age of five who present with a fever. Its complications can easily lead to acquired heart disease in children. Diagnosis typically relies on the manifestation of clinical symptoms in the traditional diagnostic process. To be diagnosed with Kawasaki disease must meet four out of five typical clinical symptoms and more than five days of fever. Five clinical symptoms include erythema of the oral mucosa(strawberry tongue or fissured lips), bilateral non-suppurative bulbar conjunctivitis, cervical lymphadenopathy, edema or erythema of the hands or feet with following peeling and polymorphous skin rash. Due to its rarity and diagnostic complexity, Kawasaki disease is often overlooked in clinical settings, leading to prolonged diagnostic timelines, and subsequently increasing the risk of coronary artery aneurysms or myocarditis

**MATERIALS AND METHODS**

In this experiment, a total of 45,274 fever cases in children under the age of seven, along with 1,785 cases of Kawasaki disease in feverish children under the same age group, will be used as data to construct and test a classification machine learning model. Data is collected from Chung Gung Medical Foundation in Taiwan, including two tertiary medical centers in Linkou and Kaohsiung and two regional hospitals in Keelung and Chiayi. By employing machine learning techniques, we will utilize routine blood test data (including CBC, WBC/DC, and partial data from biochemical tests) along with age to establish an objective predictive system. After confirming the difference between the two groups using T-test, we have 26 blood test parameters as features to use for modeling. The p-value after T-test for all features between two groups are under 0.001 as threshold. The model chosen for this experiment is LightGBM, which was introduced by Microsoft in 2017.

**RESULT AND DISCUSSION**

When evaluating the model's performance on the test set, we obtained a sensitivity of approximately 0.886 and a specificity of approximately 0.921 on full features model. The area under the ROC curve (AUC-ROC) and the area under the precision-recall curve (AUC-PR) are approximately 0.957 and 0.724, respectively, as results on the test set. Also, four more models were trained based on different routine blood test prescription and one more model was trained based on the feature importance using SHAP. These models all performed over 0.75 on recall and over 0.78 on specificity. Applying calibration test on all these six models shows that the probability of results meet the reality. The generalization of models were been prove by the test set on training process. In the extra verification test of two independent cohorts comprising 142 and 12 Kawasaki disease patients, sensitivities of approximately 0.908 and 0.917, respectively, were observed on the full features model.

**CONCLUSION**

Furthermore, we have implemented the machine learning model as a web tool, which can be accessed more easily by medical and research institutions. Users have two different input options: single-case input and file input. To enhance practical usability, we have developed multiple models based on various routine blood tests, such as CBC, WBC/DC, and biochemical tests. The web tool will automatically select the appropriate model based on the input data it detects. Notably, the web tool has already been used successfully to detect patients with Kawasaki disease before the onset of clinical symptoms and provide the information including result and probability on being Kawasaki disease.

**致謝**

**目錄**

[摘要 I](#_Toc167639908)

[英文摘要 III](#_Toc167639909)

[致謝 V](#_Toc167639910)

[目錄 VI](#_Toc167639911)

[圖目錄 VIII](#_Toc167639912)

[表目錄 X](#_Toc167639913)

[研究背景與動機 1](#_Toc167639914)

[1.1 趨勢理論介紹 1](#_Toc167639915)

[1.2 常見趨勢及趨勢反轉點定義介紹 3](#_Toc167639916)

[1.3 文獻回顧 4](#_Toc167639917)

[1.4 研究動機與目標 5](#_Toc167639918)

[第二章 資料分析與實驗流程 6](#_Toc167639919)

[2.1 研究資料 6](#_Toc167639920)

[2.2 研究流程 7](#_Toc167639921)

[2.3 資料前處理介紹 8](#_Toc167639922)

[2.4 模型介紹 12](#_Toc167639923)

[2.5 預測結果後處理介紹 14](#_Toc167639924)

[2.6 評分指標 17](#_Toc167639925)

[第三章 結果與討論 26](#_Toc167639926)

[3.1 全特徵模型 26](#_Toc167639927)

[3.2 血液常規檢查數據處方分析 33](#_Toc167639928)

[3.3 SHAP特徵重要性分析 39](#_Toc167639929)

[3.4 獨立測試資料 44](#_Toc167639930)

[3.5 與臨床特徵關聯性分析 48](#_Toc167639931)

[第四章 網站架構與網頁介紹 50](#_Toc167639932)

[4.1 網站架構介紹 50](#_Toc167639936)

[4.2 KDPredictor網頁介紹 53](#_Toc167639942)

[4.3 案例探討 56](#_Toc167639943)

[4.4 Docker工具介紹 59](#_Toc167639944)

[第五章 結論與未來展望 60](#_Toc167639949)

[5.1 結論 60](#_Toc167639955)

[5.2 未來展望 60](#_Toc167639962)

[第六章 附錄 61](#_Toc167639963)

[第七章 參考文獻 62](#_Toc167639964)

**圖目錄**

圖1.1 川崎症臨床症狀示意圖[1] 2

圖1.2 川崎症臨床症狀真實圖 2

圖1.3 卡介苗紅腫 3

[圖2.1 川崎症與其餘發燒病患年齡對資料數量對照圖 7](#_Toc167217337)

[圖2.2 資料缺值處理與過濾流程圖 7](#_Toc167217338)

[圖2.3 實驗模型訓練流程圖 8](#_Toc167217339)

[圖2.4 T值與p值解釋圖 9](#_Toc167217340)

[圖2.5 常態分布與分布機率 11](#_Toc167217341)

[圖2.6 本實驗Z分數轉換流程圖 11](#_Toc167217342)

[圖2.7 取後放回過採樣示意圖 12](#_Toc167217343)

[圖2.8 交叉驗證資料切分圖以五摺驗證為例 13](#_Toc167217344)

[圖2.9 交叉驗證訓練流程以五摺驗證為例 14](#_Toc167217345)

[圖2.10 LightGBM發展史 14](#_Toc167217346)

[圖2.11 直方圖與差加速示意圖 15](#_Toc167217347)

[圖2.12 單邊梯度採樣計算範本 16](#_Toc167217348)

[圖2.13 互斥特徵綑綁計算示意圖 17](#_Toc167217349)

[圖2.14 混淆矩陣 18](#_Toc167217350)

[圖2.15 以模型損失以及F1分數作為依據之學習曲線 19](#_Toc167217351)

[圖2.16 ROC曲線示意圖 21](#_Toc167217352)

[圖2.17 PR曲線示意圖 21](#_Toc167217353)

[圖2.18 Calibration曲線示意圖 22](#_Toc167217354)

[圖2.19 Sigmoid函式校正示意圖[8] 23](#_Toc167217355)

[圖2.20 isotonic函式校正示意圖[8] 23](#_Toc167217356)

[圖2.21 Shapely value計算範例 24](#_Toc167217357)

[圖2.22 SHAP套件點狀圖、蜜蜂圖、長條圖範例 25](#_Toc167217358)

圖3.1 全特徵模型學習曲線圖 29

圖3.2 Light GBM測試集預測分布結果圖 30

圖3.3 全特徵8種模型測試集PR與ROC曲線圖 32

圖3.4 全特徵Light GBM模型Calibration檢驗圖 33

圖3.5 血液檢查處方測試集PR與ROC曲線 36

圖3.6 年齡+CBC處方Calibration驗證圖 37

圖3.7 年齡+CBC+WBC\DC處方Calibration驗證圖 37

圖3.8 年齡+CBC+WBC\DC+ALT處方Calibration驗證圖 38

圖3.9 年齡+CBC+WBC\DC+CRP處方Calibration驗證圖 38

圖3.10 SHAP分析於測試集結果 39

圖3.11 SHAP縮減模型學習曲線圖 41

圖3.12 SHAP縮減模型與全特徵模型PR與ROC曲線圖 43

圖3.13 SHAP縮減模型Calibration檢驗圖 44

圖4.1 KDPredictor 網頁工具架構圖 50

圖4.2 Django MTV架構基本流程圖 52

圖4.3 單病患輸入區域 53

圖4.4 網頁運作流程 54

圖4.5 多病患輸入區域 55

圖4.6 模型範例輸出結果 55

圖4.7 案例一臨床症狀圖 57

圖4.8 案例一驗血數值與模型預測結果 57

圖4.9 案例二臨床症狀圖 58

圖4.10 案例二驗血資料與模型預測結果 58

圖4.11 Docker基本架構展示圖 59

**表目錄**

[表2.1 特徵p-value、兩群體各自平均值與標準誤差 9](#_Toc167217530)

[表3.1 全特徵模型CBC檢驗特徵 26](#_Toc167217550)

[表3.2 全特徵模型WBC/DC檢驗特徵 26](#_Toc167217551)

[表3.3 全特徵模型生化檢驗特徵 26](#_Toc167217552)

[表3.4 五摺驗證各批次訓練與驗證資料筆數 27](#_Toc167217553)

[表3.5 模型參數設定表 28](#_Toc167217554)

[表3.6 全特徵8種模型驗證集平均比較表 29](#_Toc167217555)

[表3.7 全特徵8種模型測試集比較表 31](#_Toc167217556)

[表3.8 血液檢查處方分群驗證集平均結果表 34](#_Toc167217557)

[表3.9 血液檢查處方分群測試集結果表 35](#_Toc167217558)

[表3.10 SHAP縮減模型驗證集平均結果表 42](#_Toc167217559)

[表3.11 SHAP縮減模型測試集結果表 42](#_Toc167217560)

[表3.12 奇美醫院獨立測試資料結果表 46](#_Toc167217561)

[表3.13 荊州第一人民醫院混淆矩陣表 47](#_Toc167217562)

[表3.14 荊州第一人民醫院評分參數表 47](#_Toc167217563)

[表3.15 OLS回歸模型結果表 49](#_Toc167217564)

1. 研究背景與動機

本章節將概述背景知識和相關現有文獻的探討。1.1節將介紹標普500及趨勢理論，1.2節將說明常見的趨勢和趨勢轉折點的定義，1.3節將回顧現有文獻中所採用的預測方法，1.4節將總結前述內容，並基於此提出本研究的動機和目標。

* 1. 背景知識介紹
     1. 標普500介紹

　　標準普爾500指數（S&P 500，簡稱標普500）是美國股票市場中最具代表性的一個股票指數之一，由普爾道瓊斯指數公司編制。其成分股涵蓋了500家在美國主要證券交易所上市的大型企業，包括科技、金融、醫療保健、能源、消費品、工業等多個領域。標普500指數被視為美國經濟和股市的重要指標，能夠提供關於經濟健康狀況和市場趨勢的廣泛視角。

　　由於其廣泛的代表性和權威性，標普500指數被廣泛用作基準指數，許多投資產品（如指數基金和交易所交易基金）都以其為標的。此外，分析師和投資者經常使用標普500指數來衡量市場趨勢和制定投資策略。

　　標普500指數不僅是美國股市的關鍵指標，還在全球範圍內具有深遠的影響力。它的表現不僅影響投資者的信心，也影響著經濟政策的制定和企業決策的方向。

* + 1. 趨勢理論介紹

　　趨勢理論是金融市場分析中的一個重要概念，旨在透過觀察市場價格的運動方向，來預測未來的價格變動。趨勢理論的核心在於假設市場價格會在一定的方向上持續運動，直到出現反轉信號。這一理論被廣泛應用於技術分析中，並成為投資決策的重要依據。

　　趨勢理論主要分為兩種方面的應用，分別是趨勢預測及趨勢反轉點預測。趨勢預測關注於分析市場中形成的主要趨勢，如上升趨勢或下降趨勢，以預測未來價格走勢的可能方向，並跟隨趨勢制定交易策略，趨勢走向與歷史資料較為相關，預測難度較低。趨勢反轉點的預測則關注於分析趨勢的變化，包括識別趨勢可能結束並轉向反方向的時機，趨勢轉折點的出現通常難以透過歷史資料判斷，因此預測難度較高。

* 1. 常見趨勢及趨勢反轉點定義介紹
     1. 常見趨勢定義介紹

　　常見的趨勢定義可以基於不同的時間跨度進行分類：

1. 以隔日收盤價較今日收盤價為基準：

若隔日的收盤價高於今日的收盤價，則被視為上升趨勢。若隔日的收盤價低於今日的收盤價，則被視為下降趨勢。以隔日漲跌做為趨勢定義，雖然預測難度低，但容易受短期波動影響，無法建立長期之交易策略。

1. 以未來數日後收盤價較今日收盤價為基準：

若未來數日後的收盤價高於今日的收盤價，則被視為上升趨勢。若未來數日後的收盤價低於今日的收盤價，則被視為下降趨勢。以未來數日後漲跌做為趨勢定義，雖然預測難度低，但難以即時預測反轉點的出現，容易誤判未來趨勢的走勢而發生虧損。

* + 1. 常見趨勢反轉點定義介紹

常見的趨勢反轉點定義有以下三種分類:

1. 以局部極值為定義

　　反轉點定義為在一定範圍內收盤價的局部極大值和極小值。當某日的收盤價是前後X日內的最高或最低點時，該日即被視為趨勢反轉點。

1. 以漲跌反轉為定義

　　收盤價連續下跌 N 天並第T日在之後連續上漲 M 天定義為趨勢向上反轉，T為趨勢反轉點:

或連續上漲 N 天並在第T日之後連續下跌 M 天定義為向下趨勢反轉，T為趨勢反轉點:

　　但市場中很少出現兩天以上的連續上漲或連續下跌，因此條件常會放寬為：

在T日之前的第n日收盤價相較於T日的收盤價下跌了至少a%;

且在T日之後的第m日收盤價相較於T日的收盤價上漲了至少b%。

T日即為反轉點。

或在T日之前的第n日收盤價相較於T日的收盤價上漲了至少a%;

且在T日之後的第m日收盤價相較於T日的收盤價下跌了至少b%。

T日即為反轉點。

1. 以Piecewise Linear Regression近似股價線段為定義

將股價時間序列遞歸地分割成多段，直到每段內的最大誤差不超過預設的閾值𝛿，步驟如下:

1. 初始化：自收盤價給定時間序列。  
   設置初始段為整個序列，即,。
2. 分割判斷：對於每一段 ,，計算段內數據點到段端點之間的線性插值直線的最大誤差。
3. 使用以下線性插值公式計算段內每個點的插值:
4. 計算每個點的誤差
5. 找到最大誤差
6. 遞歸分割：如果最大誤差𝐸大於或等於閾值𝛿，則找到最大誤差點E。   
   將當前段分為兩段： ,和,。
7. 對這兩個新段重複步驟2，直到每段的最大誤差不超過閾值𝛿。
8. 停止條件：當所有段的最大誤差都小於或等於𝛿時，停止分割。
9. 端點即為趨勢反轉點。

　　以漲跌反轉為定義容易同時定義多個相鄰同方向的反轉點，以Piecewise Linear Regression近似股價線段為定義與局部極值定義相近，但計算量較大。

* 1. 文獻回顧

存文獻當中包含多種用以預測川崎症之模型，其中用以預測川崎症之特徵十分多元，包含使用micro-RNA為基準作為生物標誌物為特徵[3]、以CNN模型搭配典型臨床症狀圖像作為特徵[4]、以臨床症狀有無搭配實驗室常規血液檢查資料作為特徵[5,6]、以實驗室常規血液檢查資料搭配尿液檢查資料作為特徵[7]等方式建立模型。

然而於上述各類模型當中，在使用生物標誌物情況下，其目標並非為早期檢驗，而是基於其川崎症背景理論所做之研究，而若是使用臨床特徵作為特徵，則病患應已發病數日，與使用傳統診斷方式所需時間相差不遠，因此這類的特徵並無法滿足臨床上的早期檢驗需求，無法提供醫師於發燒初期對於病患是否罹患川崎症的資訊。而尿液檢驗則非常規檢驗中會執行之測驗，因此以其作為特徵之模型，在臨床應用上能使用的時機點則較少，無法提供醫師在臨床上的泛用。故本次實驗希望能改善上述做法，提供一個具有泛用性且能提供患者早期診斷資訊之工具。

* 1. 研究動機與目標

目前對於趨勢反轉點的研究多注重於預測隔日是否為反轉點，容易發生正負樣本不平衡的問題，加上反轉點預測原本就較為困難，訓練準確的預測模型非常困難。本研究旨在透過預測未來數日間的每日趨勢，並從趨勢變化識別趨勢反轉點出現的時機。

第二章 資料與實驗流程

在本章節中，將依次介紹本次實驗所使用的資料來源以及資料分布，簡述本次實驗的流程與架構，章節2.3與2.4節將依照流程與架構內所使用到的技術做細節介紹，章節2.5將介紹本次所使用的模型以及其特點，章節2.6將介紹於機器學習模型評估上所使用的參數，以及簡述參數於統計上所具備的真實意義。

* 1. 實驗流程

本次研究資料來自於長庚醫療機構，包含長庚於林口以及高雄的三級醫療機構，與兩個座落於基隆還有嘉義的地區性醫院，總共包含超

* 1. 資料前處理

在本次實驗流程中，會先進行資料前處理，包含特徵篩選、缺值處理以及年齡過濾，當中缺值處理與

* + 1. 定義趨勢反轉點及趨勢

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* + 1. 設計輸入特徵

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* + 1. 清理資料

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* + 1. 以滑動視窗切分資料

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* + 1. 去除多餘的趨勢反轉

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* + 1. 將特徵正規化

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，

* 1. 模型訓練與預測
     1. 訓練流程

於值小於0.001則代表其拒絕虛無假說，滿足於對立假說，進一步認定該特徵於兩母體間存在顯著差異，應可作為模型特徵使用。

* + 1. 模型介紹

在使用篩選出之特徵進行模型訓練前，由於每個特徵於收集實所使用的單位有所不同，因分

* 1. 預測結果後處理

ent Boosting Decision Tree)，在此演算法當中，每次的計算皆希望能夠減少上一次計算的殘差(residual)，並以消除殘差為目的產生新模型，使用梯度作為取代權重的算法，而LightGBM又更進一步採用單邊梯度採樣(2.5.3節)減少運算所需資源

* + 1. 識別趨勢反轉點

在直方圖演算法中會將特徵數值以直方圖的方式做分類，以減少運算所需時間，將浮點數歸類於特定直方圖區間內，同時於建立完直方圖，於決策樹分枝計算時，可依靠父節點直方圖與單一側子節點直方圖計算另一節點數值分布，進行差加速，減少模型運算過程中所需時間

* + 1. 設計交易策略

單邊梯度採樣(Gradient-based One-Side Sampling, GOSS)，相較於其餘Boosting演算法而言，在LightGBM會採用以梯度值取代權重值的方式，減少運算量，於梯

* 1. 模型評估

ROC&PR曲P則用以解析模型與特徵間的關聯。

* + 1. 趨勢預測結果評估

本次實驗中為了檢

* + 1. 趨勢反轉點預測結果評估

本次實驗中為了檢

* + 1. 交易策略回測結果評估

在模型效能評估上，除

第三章 結果與討論

在本章節中將依次介紹本次實驗所得到的不同模型以及其經過模型評分參數後所得的結果以及網頁呈現，分別為3.1章節的全特徵模型，包含CBC檢驗、WBC/DC檢驗以及生化檢驗，3.2章節的依檢驗處方分群建立模型，包含第一種單純依照CBC，第二種使用包含CBC以及WBC/DC，以及第三、四種使用包含CBC、WBC/DC以及生化中的CRP或ALT此四種不同處方以及全特徵模型比較，3.3章節將依照SHAP分群做縮減模型比較，3.4章節將呈現當使用獨立資料集時，模型的效能比較，3.5章節將比較臨床症狀與模型的關聯性，3.6章將做現有網頁工具的簡易介紹。

* 1. 全特徵模型

於全特徵模型當中，本次實驗共使用26個特徵，當中包含1個獨立特徵年齡，10個來自CBC檢驗的特徵，13個來自WBC/DC檢驗的特徵以及2個來自生化檢驗的特徵。

表3.1 全特徵模型CBC檢驗特徵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WBC(1000) | Hemoglobin | HB z score | Hematocrit | MCH |
| RBC | Platelets | PLT z score | MCV | MCHC |

表3.2 全特徵模型WBC/DC檢驗特徵

|  |  |  |  |
| --- | --- | --- | --- |
| Lym(%) | Seg(%) | Eosinophil (%) | Basophil (%) |
| Lym (count) | Seg(count) | Eosinophil (count) | Basophil (count) |
| Band (%) | Band (count) | Monocyte (%) | Monocyte (count) |
| EOS z score |  |  |  |

表3.3 全特徵模型生化檢驗特徵

|  |  |
| --- | --- |
| CRP | ALT/GPT |

於章節2.1中已描述本次實驗所使用的資料總數，於本次實驗流程中將該總數資料以9:1的分布，將資料切分為訓練-驗證集合以及測試集，於訓練流程中將只使用訓練-驗證集合，測試集合將於完成參數調教後做為模型最終效能評估使用。

於本次的實驗流程中，將使用五摺驗證做為交叉驗證方式，於模型評估上，在學習曲線上會以五摺中的最大最小值做為上下限標註依據，並以實線標註其平均值，於PR與ROC曲線上同理，以最大最小值做為上下限依據以實線標註其平均值，於評分參數上採用五摺平均做為表示。

表3.4 五摺驗證各批次訓練與驗證資料筆數

|  |  |  |  |
| --- | --- | --- | --- |
| fold 1 | FC | KD | 正負比 |
| train | 32597 | 1287 | 3.95% |
| validation | 8155 | 322 | 3.95% |

|  |  |  |  |
| --- | --- | --- | --- |
| fold 2 | FC | KD | 正負比 |
| train | 32565 | 1300 | 3.99% |
| validation | 8187 | 309 | 3.77% |

|  |  |  |  |
| --- | --- | --- | --- |
| fold 3 | FC | KD | 正負比 |
| train | 32616 | 1295 | 3.97% |
| validation | 8136 | 314 | 3.86% |

|  |  |  |  |
| --- | --- | --- | --- |
| fold 4 | FC | KD | 正負比 |
| train | 32584 | 1285 | 3.94% |
| validation | 8168 | 324 | 3.96% |

|  |  |  |  |
| --- | --- | --- | --- |
| fold 5 | FC | KD | 正負比 |
| train | 32646 | 1269 | 3.89% |
| validation | 8106 | 340 | 4.19% |

於訓練流程上本次實驗中所採用的Light GBM參數為使用GBDT做為Boosting方式，並設置學習率為0.2、深度5，單層最大枝葉展開為5且同時每個枝葉內數據量不得小於10筆，當中枝葉與深度的設定為避免模型產生過擬合的問題。

表3.5 模型參數設定表

|  |  |
| --- | --- |
| 模型參數設定(LGBM) | |
| Boosting | gbdt |
| learning\_rate | 0.2 |
| Min data in leaf | 10 |
| Max depth | 5 |
| num\_leaves | 5 |

於該參數設定下，模型的學習曲線能夠有效收斂，於過採樣的訓練曲線與過採樣後的驗證曲線間，兩線段表現差異約為4~5%，且兩線段最終呈平行呈現，同時於原分布的驗證集曲線下可發現，其表現約落於F1分數接近0.43，損失約為0.08。

一張含有 文字, 螢幕擷取畫面, 陳列, 行 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 行, 繪圖 的圖片

自動產生的描述

圖3.1 全特徵模型學習曲線圖

我們將同樣資料集、流程以及模型檢驗標準套用於其餘7個經典的機器學習分類模型，分別為AdaBoost、XGBoost、Random Forest、Logistic Regression、LDA、SVM以及MLP，以比較不同模型間的效能以及優劣，於結果表上可看出在使用Boosting三種演算法以及MLP的情況下於F1分數以及PR曲線下面積表現相似，然而若考量到實際運算所需時間以及硬體成本，於使用AdaBoost約需2.5個小時、XGBoost約需10分鐘、MLP約需0.5個小時，而Light GBM僅約需2分鐘的情況下，Light GBM模型最終仍選為本次實驗所使用的最終選擇，而於線性分類器的兩方法Logistic Regression以及LDA的情況上，儘管已將模型可調控參數調整至性能最佳的情況下，仍發生欠擬合的情況，因此可猜測在本次實驗所訓練的情況以及特徵下，簡單的線性分類無法很好的區分出兩類別的不同，因此需使用維度更高的模型作為應用模型。

表3.6 全特徵8種模型驗證集平均比較表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| origin | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| Ada Boost | 0.441 | 0.296 | 0.862 | 0.919 | 0.952 | 0.630 | 0.994 | 10.689 | 0.151 |
| Light GBM | 0.451 | 0.305 | 0.868 | 0.922 | 0.956 | 0.645 | 0.994 | 11.116 | 0.144 |
| XG Boost | 0.421 | 0.277 | 0.880 | 0.909 | 0.956 | 0.635 | 0.995 | 9.717 | 0.132 |
| Random Forest | 0.398 | 0.259 | 0.865 | 0.902 | 0.946 | 0.585 | 0.994 | 8.835 | 0.150 |
| Logistic | 0.389 | 0.252 | 0.854 | 0.900 | 0.942 | 0.492 | 0.994 | 8.548 | 0.163 |
| LDA | 0.396 | 0.261 | 0.820 | 0.908 | 0.936 | 0.489 | 0.992 | 8.951 | 0.198 |
| SVM | 0.447 | 0.303 | 0.847 | 0.923 | 0.948 | 0.607 | 0.994 | 11.045 | 0.166 |
| MLP | 0.445 | 0.300 | 0.863 | 0.920 | 0.956 | 0.655 | 0.994 | 10.864 | 0.149 |

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

圖3.2 Light GBM測試集預測分布結果圖

若以五摺中PR曲線下面積表現最佳的模型作為最終模型，並將測試集套用於上，則可得於測試集測試情況下的模型效能，從中我們可看出，模型具有十分良好的泛化性，儘管在從未看過的資料情況下，模型仍有接近甚至超越驗證集平均的效能，考量到於臨床運用時，招回病人回診為一個十分重要的環節，因此於臨床應用上更著重於Recall數值的優劣，而本次全特徵模型於Recall上有接近88.6%的準確度，顯現其應可於一定程度上應用於真實情況，而在當中又以Light GBM模型於PR區域下面積表現最佳，同時由PR曲線以及ROC曲線做觀察，發現模型可明顯區分成兩群，分別為線性模型以及非線性模型，當中線性模型於PR曲線上面積會發生朝向原點收斂的現象，其應源自模型的欠擬合，此外於圖中可明顯看出Boosting三種演算法呈現表現較佳的趨勢。

表3.7 全特徵8種模型測試集比較表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| Ada Boost | 0.439 | 0.293 | 0.875 | 0.918 | 0.952 | 0.699 | 0.995 | 10.636 | 0.136 |
| Light GBM | 0.452 | 0.303 | 0.886 | 0.921 | 0.957 | 0.724 | 0.995 | 11.165 | 0.123 |
| XG Boost | 0.419 | 0.273 | 0.898 | 0.907 | 0.955 | 0.710 | 0.996 | 9.643 | 0.113 |
| Random Forest | 0.405 | 0.262 | 0.898 | 0.901 | 0.954 | 0.675 | 0.996 | 9.102 | 0.113 |
| Logistic | 0.388 | 0.249 | 0.875 | 0.897 | 0.937 | 0.475 | 0.995 | 8.527 | 0.139 |
| LDA | 0.423 | 0.280 | 0.869 | 0.913 | 0.936 | 0.471 | 0.994 | 9.977 | 0.143 |
| SVM | 0.469 | 0.320 | 0.881 | 0.927 | 0.951 | 0.660 | 0.995 | 12.068 | 0.128 |
| MLP | 0.475 | 0.324 | 0.886 | 0.928 | 0.955 | 0.697 | 0.995 | 12.333 | 0.122 |

一張含有 文字, 螢幕擷取畫面, 繪圖, 圖表 的圖片

自動產生的描述一張含有 文字, 行, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

RECALL

圖3.3 全特徵8種模型測試集PR與ROC曲線圖

為了驗證全特徵Light GBM模型的結果機率可靠性，本次實驗採用Calibration方式做模型檢驗，於Calibration檢驗流程時將對FC資料部分做降採樣，以滿足資料比，於以此方式所產生的檢驗圖上可看出，模型整體趨勢接近對角完美機率，因此可視為其機率值具有一定可信度，並不會過度樂觀呈現，而藉由分布圖以及回歸方式同樣，可看出本次模型提供的預測機率值多數坐落於0以及1附近，靠近0.5附近的機率值分布較少。

一張含有 文字, 圖表, 行, 繪圖 的圖片

自動產生的描述

圖3.4 全特徵Light GBM模型Calibration檢驗圖

* 1. 血液常規檢查數據處方分析

雖然全特徵模型的效能表現卓越，且其優秀泛化性應該也能夠讓該模型於實際情況上應用，然而於真實情況上，並非每間醫療機構皆會在接受到病人的時候便開立三種血液檢查處方，部分機構可能僅開立CBC或是CBC與WBC/DC，同時也並非每所醫療機構皆有能力執行三種不同的血液檢查，此外儘管該醫療機構同時開立三種血液檢查處方，於生化檢驗內也不能保證會同時具有CRP以及ALT/GPT兩欄位，同時也可能於資料抄寫、挪移過程中導致特徵數值的缺失，因此為了因應多種情況，以及使得本次實驗所建立的模型能夠更符合真實應用情況，本次實驗會再依照血液檢查處方將模型分成下列五種情況：

* 1. 年齡 + CBC +WBC/DC + 生化 (全特徵)
  2. 年齡 + CBC
  3. 年齡 + CBC + WBC/DC
  4. 年齡 + CBC + WBC/DC + ALT
  5. 年齡 + CBC + WBC/DC + CRP

上述五種模型分類中，除去全特徵模型已於3.1章中提及過外，其餘模型將以同全特徵模型的五摺驗證資料，將特徵刪減至符合該情況後重新訓練模型，由於已於3.1章證明Light GBM於模型效能綜合表現上較為優異，故於此額外四種模型將僅使用Light GBM作為訓練模型，並同樣以相同類型的學習曲線確保模型的擬合情況，避免模型發生過擬合或欠擬合，以此方式所得驗證集平均表現如下表3.8。

表3.8 血液檢查處方分群驗證集平均結果表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| origin | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| 全特徵 | 0.451 | 0.305 | 0.868 | 0.922 | 0.956 | 0.645 | 0.994 | 11.116 | 0.144 |
| CBC | 0.201 | 0.116 | 0.741 | 0.778 | 0.837 | 0.234 | 0.987 | 3.339 | 0.333 |
| CBC+  DC | 0.315 | 0.196 | 0.792 | 0.872 | 0.912 | 0.472 | 0.991 | 6.193 | 0.238 |
| CBC+  DC+ALT | 0.376 | 0.243 | 0.832 | 0.898 | 0.936 | 0.578 | 0.993 | 8.136 | 0.188 |
| CBC  +DC+CRP | 0.399 | 0.260 | 0.856 | 0.904 | 0.945 | 0.582 | 0.994 | 8.895 | 0.159 |

於表3.8當中可看出，隨著使用的單據增加，模型於F1分數以及Recall上有顯著的提升，因此可看出三張處方皆使用的全特徵模型存在必要性，同時雖然使用較少特徵時，模型的效能會有所下降，但於僅用CBC的情況下，驗證集平均Recall仍有74.1%的準確度，於臨床上作為參考工具已為一個十分強力的工具，而為了檢驗不同模型的泛化性，同樣使用測試集作為檢驗資料，當中將只採用個模型訓練所需特徵，測試集結果表如表3.9。

表3.9 血液檢查處方分群測試集結果表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| 全特徵 | 0.452 | 0.303 | 0.886 | 0.921 | 0.957 | 0.724 | 0.995 | 11.165 | 0.123 |
| CBC | 0.214 | 0.124 | 0.778 | 0.786 | 0.855 | 0.277 | 0.989 | 3.633 | 0.282 |
| CBC+  DC | 0.309 | 0.190 | 0.835 | 0.861 | 0.915 | 0.511 | 0.993 | 6.014 | 0.191 |
| CBC+  DC+ALT | 0.367 | 0.234 | 0.847 | 0.892 | 0.934 | 0.641 | 0.993 | 7.861 | 0.172 |
| CBC  +DC+CRP | 0.410 | 0.267 | 0.886 | 0.905 | 0.949 | 0.632 | 0.995 | 9.343 | 0.126 |

測試集選用模型邏輯同全特徵模型，將從五摺驗證中PR曲線下表現最佳之模型作為最終模型，並將測試集套用於上得到測試集結果表，於表中可看出模型於測試集上同樣顯現出隨特徵增加，效能愈加的趨勢，然僅使用CBC之模型於測試集上也有近77.8%的Recall準確度，因此可看出模型具有一定泛化性，因此推測模型應也可適用於不同的資料集上，進一步應用於臨床實際應用。

一張含有 文字, 圖表, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

一張含有 文字, 行, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

圖3.5 血液檢查處方測試集PR與ROC曲線

於PR以及ROC曲線圖上，我們可更直觀看出每個模型的魯棒性，於PR圖上表現顯示當減少模型特徵數或減少處方單據時，曲線明顯朝向原點且區域下面積較低，呈現出效能較低的情況，而當特徵與處方單據增加時，則明顯朝向(1,1)方向靠近，各模型間有明顯的差異性，此現象同樣發生於ROC曲線上。

一張含有 文字, 圖表, 螢幕擷取畫面, 行 的圖片

自動產生的描述

圖3.6 年齡+CBC處方Calibration驗證圖

一張含有 文字, 圖表, 行, 地圖 的圖片

自動產生的描述

圖3.7 年齡+CBC+WBC\DC處方Calibration驗證圖

一張含有 圖表, 行, 文字, 方案 的圖片

自動產生的描述

圖3.8 年齡+CBC+WBC\DC+ALT處方Calibration驗證圖

一張含有 圖表, 文字, 行, 方案 的圖片

自動產生的描述

圖3.9 年齡+CBC+WBC\DC+CRP處方Calibration驗證圖

為了檢驗模型預測機率得可信度，在此將除全特徵模型外四種驗血處方模型以降採樣其餘發燒病患方式座Calibration驗證，由Calibration驗證圖我們可看出此四種不同驗血處方單據分群結果皆接近完美機率線，故其預測性質應是具有一定可信度，並無明顯的過度樂觀，若進一步分析觀察每種單據分群間的圖形差別，我們可看出隨著使用單據逐漸增加，模型對於預測結果的分布會逐漸兩極化，代表模型能夠給出非常肯定為川崎症或非常肯定為發燒病患的預測結果，此現象同樣會顯示在模型效能上，由ROC圖上可看出單只用CBC抽血單據作為建模特徵時，其效能表現較差，與Calibration上分布平均相呼應，此兩者關連同樣隨抽血單據分群不同有所改變，故Calibration與ROC間的關聯與結果相符。

* 1. SHAP特徵重要性分析

為了進一步解釋全特徵模型內各特徵對於模型的貢獻度多寡，本次實作採用SHAP套件作為分析工具，該套件所使用依據為Shapley Value作為評估依據，該數值計算原理基於賽局理論，詳細計算公式與其餘介紹已於章節2.6.5中提及，該次實作由於Light GBM為決策樹為基底所開發出的Boosting模型，故在此將使用TreeExplainer計算Shapley Value，將測試集套用於該套件上，並統計對於模型重要程度前十名的特徵，該十名特徵分別如下圖。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

圖3.10 SHAP分析於測試集結果

本次實驗中，為了進一步檢驗其分析結果以及建造一個更輕量化的模型，將採用分析結果圖中，對於模型貢獻程度前五名的特徵建立一個縮減模型，所使用特徵如下:

* 1. CRP
  2. ALT/GPT
  3. EOS z score
  4. Monocyte (%)
  5. PLT z score

於建立流程內，將同樣採用五摺驗證作為訓練與模型調教方式，過程中各摺之間所使用資料同全特徵模型，僅做特徵篩選，並依照學習曲線作為模型調教依據，並以五摺平均評分參數作為驗證集整體效能評估，於中選擇效能最佳摺作為最終選擇模型，並以最終選擇模型做測試集評估。

一張含有 文字, 螢幕擷取畫面, 陳列, 行 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 行, 繪圖 的圖片

自動產生的描述

圖3.11 SHAP縮減模型學習曲線圖

於學習曲線圖上可看出，模型最終收斂且於F1分數上誤差約於5%內，在此若比較該縮減模型與全特徵模型的學習曲線圖，能發現兩者圖形走勢類似，且F1分數最終收斂分數接近，而於驗證集評分參數表上，可看出模型於各項參數上表現皆與

全特徵模型相似，當中Recall更僅有約0.01左右的差距，因此可推測兩模型於效能上應有接近表現，在此可能是源於SHAP分析圖上的不平均，於圖形上可看出排名第一的CRP相較其他特徵而言佔有絕對領先地位，且後續特徵重要性隨排名急速降低，於5名後的特徵其貢獻度皆為0.4之下，甚至於10名後僅為0.1之後，故僅以前五名特徵所建模型能與全特徵模型具有相似的效能表現。

表3.10 SHAP縮減模型驗證集平均結果表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| 全特徵 | 0.451 | 0.305 | 0.868 | 0.922 | 0.956 | 0.645 | 0.994 | 11.116 | 0.144 |
| Top5 | 0.406 | 0.266 | 0.858 | 0.907 | 0.948 | 0.620 | 0.994 | 9.192 | 0.157 |

為了檢驗其泛化性，將測試集應用於最終模型上可得下表3.10，從表中我們可看出不論在臨床上較具有識別的Recall或是PR曲線下面積，在測試集上表現皆與驗證集相似，甚至更優於驗證集，因此可視為模型具有一定的泛化程度，而若與全特徵模型相比，雖於PR曲線下面積表現上略遜於全特徵模型性能，但於Recall上則有與全特徵模型相比擬的效能，可看出若僅單純用前五名建設模型，其效能依然具有一定的可靠性。

表3.11 SHAP縮減模型測試集結果表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | **F1** | **Precision** | **Recall** | **Specificity** | **ROC** | **PR** | **NPV** | **LR +** | **LR -** |
| 全特徵 | 0.452 | 0.303 | 0.886 | 0.921 | 0.957 | 0.724 | 0.995 | 11.165 | 0.123 |
| Top5 | 0.435 | 0.288 | 0.886 | 0.915 | 0.942 | 0.695 | 0.995 | 10.384 | 0.124 |

若進一步以PR曲線圖與ROC曲線圖上做觀察，可發現SHAP前五名縮減模型與全特徵模型在圖形表現上十分接近，於PR曲線圖上，於Recall值較大時兩者近乎重合，微小差異體現於當Recall值較低時，造成區域下面積微小差異，於ROC曲線上差異於當偽陽性率大於0.4之上，使得其區域下面積造成微小差異，然總體而言趨勢與面積接近，顯示兩模型具有相似的效能呈現。

一張含有 文字, 圖表, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

圖3.12 SHAP縮減模型與全特徵模型PR與ROC曲線圖

於進行Calibration檢驗於SHAP縮減模型時，同樣將發燒病患做降採樣後進行分析，可看出其曲線接近完美機率線，可看出模型的可信度，且縮減模型的預測機率兩極化，與全特徵模型相似，預測機率多數落於1與0附近。

一張含有 圖表, 行, 方案, 文字 的圖片

自動產生的描述

圖3.13 SHAP縮減模型Calibration檢驗圖

將SHAP縮減模型所使用到的特徵作進一步的分析，在此將縮減模型特徵依照常規抽血檢查處方單據分群，可將特徵分成下列情況:

1. 屬於CBC抽血單據特徵 : PLT z score
2. 屬於WBC\DC抽血單據特徵 : EOS z score , monocyte (%)
3. 屬於生化抽血單據特徵 : CRP, ALT\GPT

從上述將縮減模型特徵依照抽血單據分群中，我們可輕易看出，該縮減模型所使用到的特徵來自於三種不同的抽血單據，且同時特徵重要性最高的前兩名皆來自於生化抽血單據，這同時符合川崎症的全身性發炎特性，ALT\GPT與CRP此兩生物指標能夠很有效反應出病人發炎狀況，當中又以CRP反應速度最快，符合該特徵於模型重要度第一名的性質，同時，分布於三種抽血單據同樣應證於3.2節中的結果，於3.2節中，我們看出隨抽血單據使用增加，模型效能也相對應提升，且當加入生化單據時最為明顯，這同樣與本節中分群相對應，前五名特徵分別落於三種不同單據，故若單使用CBC，效能最為差勁，而加入生化時較為大幅上升的效能也應證於上述所提的前兩名特徵皆為生化單據內所提供。

* 1. 獨立測試資料

儘管於訓練流程中，已將測試集單獨與訓練-驗證集分開使用，但由於其資料收集背景時空與機構相同，難免造成資料仍有一定程度的相似性，而造成無法真實模擬現實情況應用於不同機構、時空，也會造成無法正確驗證模型泛化能力，故本次實驗中收集不同獨立測試資料用以檢測模型的真實效能與泛化能力，分別為同機構但不同時間段收集之資料集、同國家不同機構所收集之資料集，以及不同國家不同機構所收集之資料集。

* + 1. 同機構不同時間段收集之資料集

不同於原始收集數據，在同機構但不同時間段收集到的資料集收集於Covid-19時間來自於長庚醫療體系的川崎症病患資料，本次所收集資料內共有142筆急性川崎症病患作為正資料集，以及4筆非急性或非川崎症病患資料作為負資料集，由於當中負資料集欠缺，故本次僅以臨床較為注重的Recall作為評分參數呈現，負資料集則以文字表述，其結果如下:

1. 全特徵模型
   1. Recall : 0.908
   2. 非急性KD或非KD病患 : 4筆中辨別出4筆
2. 年齡+CBC模型
   1. Recall : 0.739
   2. 非急性KD或非KD病患 : 4筆中辨別出2筆
3. 年齡+CBC+WBC\DC模型
   1. Recall : 0.859
   2. 非急性KD或非KD病患 : 4筆中辨別出3筆
4. 年齡+CBC+WBC\DC+CRP模型
   1. Recall : 0.915
   2. 非急性KD或非KD病患 : 4筆中辨別出3筆
5. 年齡+CBC+WBC\DC+ALT模型
   1. Recall : 0.894
   2. 非急性KD或非KD病患 : 4筆中辨別出3筆
6. SHAP前五名縮減模型
   1. Recall : 0.873
   2. 非急性KD或非KD病患 : 4筆中辨別出4筆

從結果中可看出Recall的數值與訓練流程中的驗證集與測試集接近，故可視各模型就算於不同時間段中收集資料，若出於同機構，仍有一定程度的辨識效能，非急性或非川崎症病患當中，病患包含已施打IVIG治療與MIS-C病患，於全特徵模型與SHAP縮減特徵模型中，仍可將這些與急性川崎症病患相似案例很好的辨別出來，可看出這些模型於辨別負資料集上的可信度。

* + 1. 同國家不同機構所收集之資料集

於同國家不同機構所收集之資料集，本次實驗收集來自於奇美醫院的川崎症以及非川崎症病患，當中病患IVIG治療次數皆過濾為1次，於此情況下由於所使用資料具有部分缺值，故每個模型所能使用的數量有所不同，而此缺值情況也更加符合真實情況，各模型所能使用資料數量與模型效能如下表3.12。

表3.12 奇美醫院獨立測試資料結果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型總類** | **可使用數據** | **KD預測數量** | **NonKD預測數量** | **Recall** |
| Full Model | 141 | 118 | 23 | 0.837 |
| CBC | 189 | 128 | 61 | 0.677 |
| CBC+WBC | 189 | 146 | 43 | 0.772 |
| CBC+WBC+ALT | 141 | 118 | 23 | 0.837 |
| CBC+WBC+CRP | 188 | 153 | 35 | 0.814 |
| Top5 | 141 | 118 | 23 | 0.837 |

於表3.12中可看出，隨著模型所使用特徵增加，所能使用的病患數量與之降低，其原因便來自於每個醫療機構所開立處方單據不同以及收集資料時謄寫缺值所導致，故提供多模型給予使用者便成為一個重要的功能，於Recall表現上各模型表現性能略遜於訓練流程內的驗證集平均與測試集，然其差異不大，於資料整理環節中，由於不同機構對於年齡欄位所收集有少部分不同，故於以年齡作為z score依據時，可能造成些微差距導致部分特徵精確性上有些微疑慮，然而除去僅使用CBC模型外，其餘模型正確率仍有接近80%表現，各模型間對於每個病患的結果預測也能夠提供醫師更多方面的考量。

* + 1. 不同國家不同機構所收集之資料集

於不同國家不同機構所收集之資料中，本次實驗收集來自中國荊州人民第一醫院的川崎症以及非川崎症資料病患分別做為正資料集以及負資料集，於此資料集中，部分資料於特定特徵中含有缺值的存在，因此於模型測試階段，每個模型所能使用的病患數量不同，會依照各模型所能使用到的病患分開預測，在此情況下的預測結果混淆矩陣如下表3.12所示，評分參數如下表3.13所示。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model case** | **model used cases** | **real KD (used)** | **real FC (used)** | **true positive** | **false positive** | **true negative** | **false negative** |
| 全特徵模型 | 20 | 12 | 8 | 11 | 1 | 7 | 1 |
| CBC | 178 | 32 | 146 | 28 | 54 | 92 | 4 |
| CBC+WBC | 23 | 15 | 8 | 14 | 2 | 6 | 1 |
| CBC+WBC+ALT | 21 | 13 | 8 | 11 | 2 | 6 | 2 |
| CBC+WBC+CRP | 21 | 13 | 8 | 11 | 1 | 7 | 2 |
| Top5 | 173 | 25 | 148 | 21 | 6 | 142 | 4 |

表3.13 荊州第一人民醫院混淆矩陣表

表3.14 荊州第一人民醫院評分參數表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model case** | **recall** | **precision** | **specificity** | **accuracy** | **NPV** | **P\_LR** | **N\_LR** |
| Full model | 0.916667 | 0.916667 | 0.875 | 0.9 | 0.875 | 7.333333 | 0.095238 |
| CBC | 0.875 | 0.341463 | 0.630137 | 0.674157 | 0.958333 | 2.365741 | 0.19837 |
| CBC+WBC | 0.933333 | 0.875 | 0.75 | 0.869565 | 0.857143 | 3.733333 | 0.088889 |
| CBC+WBC+ALT | 0.846154 | 0.846154 | 0.75 | 0.809524 | 0.75 | 3.384615 | 0.205128 |
| CBC+WBC+CRP | 0.846154 | 0.916667 | 0.875 | 0.857143 | 0.777778 | 6.769231 | 0.175824 |
| Top5 | 0.84 | 0.777778 | 0.959459 | 0.942197 | 0.972603 | 20.72 | 0.166761 |

由上表3.13 中，可看出當今天缺值嚴重時，全特徵模型所能應用的病例較少，相比而言僅使用CBC作為特徵所建立出的模型仍有178個病患能夠使用，因此進一步驗證出使用不同驗血處方單據作為可應用模型的重要性。

從表3.14中可看出，在全特徵模型的情況下，模型的Recall值仍有約91.67%，顯示儘管今天的資料來源是源自不同的機構甚至是不同國家的情況上，只要提供的模型特徵單位與原先要求相同，模型仍有一定泛化性，能夠很好的辨別出川崎症病患。在同表3.13中，若僅觀察使用病患最多而原先模型效能最差的僅使用CBC模型，在這次的資料集中，仍具有87.5%的Recall值，代表在縮減特徵的情況下，模型仍然可以具有很好的表現性能，同時該模型又可使用於較多病歷上，進一步驗證該模型的臨床應用可行性。

* 1. 與臨床特徵關聯性分析

於本次實驗中將使用OLS分析將以臨床症狀作為特徵並以案例預測機率結果作為回歸目標。

* + 1. OLS分析

普通最小平方法(Ordinary Least Square, OLS)，為線性回歸中的一種計算方式，以殘差平方和(9)最小作為目標，用以估計母體參數。

(9)

回歸結果會依照R平方值(10)以及F檢定的p-value作為模型評判標準，當R平方值愈大時，則代表模型有較佳的解釋力，能夠將特徵與回歸目標間關係做連結，反之則否。

(10)

F檢定於OLS回歸中，可用以鑑定回歸方程式中斜率是否接近於0，進一步驗證回歸關係中的顯著性，當p-value愈小，則代表回歸關係中相較顯著，同樣將用於本次作為OLS回歸中關聯性分析的一個評分指標。

* + 1. OLS分析結果

本次實驗中，將以病患臨床症狀作為OLS回歸模型特徵，並以各病患使用模型預測機率結果為回歸目標，當中臨床特徵會依照症狀強度以及有無，分為0、1、2三種數值，本次所使用特徵共七個，包含如下:

1. 嘴唇病變
2. 眼部病變
3. lymph>1.5
4. lymph腫脹
5. 水腫
6. 皮膚紅斑
7. 卡介苗接種處紅腫

本次使用含臨床症狀資料共包含677個川崎症病患，然而並非所有病患皆具有全部特徵能夠使用每種模型，故下表3.14中將提供各模型所能使用病患數量，以及OLS回歸模型使用能預測病患下所得出的R平方值與F檢定p-value。

表3.15 OLS回歸模型結果表

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **使用病患數** | **R平方值** | **F檢定p-value** |
| 全特徵 | 623 | 0.05 | 4.15e-5 |
| 年齡+CBC | 677 | 0.026 | 0.014 |
| 年齡+CBC+WBC\DC | 675 | 0.038 | 5.49e-4 |
| 年齡+CBC+WBC\DC+ALT | 631 | 0.04 | 6.811e-4 |
| 年齡+CBC+WBC\DC+CRP | 665 | 0.034 | 1.89e-3 |
| SHAP 前五名 | 623 | 0.039 | 9.57e-4 |

由表3.15中可看出，儘管F檢定的p-value多數模型中皆小於常見的分類閾值0.001或0.0001，但由於R平方值於所有模型情況上皆趨近於0，表示模型斜率較小，顯現特徵與預測機率間無明顯關聯性，此結果進一步顯示使用機器學習模型作為診斷輔助工具的重要以及方便性，機器學習能夠不仰賴臨床症狀，從單純客觀數值角度給予醫師不同的參考依據，同時具有一定的可信效能。

第四章 網站架構與網頁介紹

3. 1. 網站架構介紹

本次實驗中，除了單純的機器學習效能展現，以及不同統計方法上的驗證，為了使醫師能夠於臨床上實際投入應用，因此與來自長庚醫療機構的川崎症專家郭和昌醫師合作，在與目前川崎症發表文章排名第一，並多年深耕於該領域的郭和昌醫師合作下共同開發出由python Django套件作為背景，提供多種常規驗血處方單據模型與SHAP分析模型的網頁工具，同時除使用伺服器架設之網站外，同時也使用Docker作為開發環境，以提供本地端運行，方便偏遠醫療機構於網路或通訊不方便時，仍可有一個穩定使用的網頁工具，下圖4.1為本次網站架構。

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

圖4.1 KDPredictor 網頁工具架構圖

* + 1. 網頁前端介紹

於網頁前端本次使用傳統的HTML、CSS作為靜態網頁編成，於HTML中使用Bootstrap作為網頁美觀以及排版套件，使用Bootstrap能夠確保網頁外觀的一致性，同時也能夠簡化開發流程，於動態網頁呈現以及與後端的響應上使用JavaScript作為編成語法，並使用jQuery作為部分API，在網頁動態處理上使用AJAX方式與後端做溝通，AJAX為一種非同步的網頁前後端溝通方式，於AJAX技術中除了使用非同步之外，同時也使用token作為溝通依據，以此能夠確保資料的安全性，於表格呈現上使用DataTable套件，能夠讓使用者有一個更好的表格使用體驗，且簡化於表格設計上所需的開發時間。

* + 1. 網頁後端介紹

Django為一個以python作為網頁後端，並以HTML、JavaScript、CSS作為網頁前端所設立的網頁開發基礎，Django架構將開發流程模組化，使得後端能夠輕易整合並呼叫不同python套件，且基於現階段機器學習開發流程多使用python套件，因此使用Django作為網頁後端，能夠教輕易將模型與網站做串接。

Django繼承傳統的MCV架構並進一步將其改良成專屬的MTV架構，MTV架構中包含Model、Template、View三部分。

Model部分主要為資料模組，用以與資料庫做串接並處理，於Django架構中，使用Object Relational Mapping(ORM)語法作為與資料庫溝通的API，該強大的API能夠使開發人員以簡潔方便的方式達到與資料庫做串接， ORM會將底層資料庫抽象化為Model，使得不同資料庫能夠無縫做溝通與串接，並且ORM語法具有更高的安全性，能夠防止常見的攻擊，例如 SQL injection，儘管如此Django也支援傳統SQL語法，開發人員也能夠以傳統SQL語法與資料庫做大量資料且快速的溝通，

Template主要為視圖模組，即是傳統的網頁呈現部分，當中多存放網頁靜態呈現所需HTML模板，Template當中所呈現的靜態內容會由控制模組的HTTP回應後再進行渲染，如此便能夠將後端運算結果於前端做渲染呈現。

而View為控制模組，為主要後段運算所使用部分，當使用者於前端發送HTTP要求後，便會依照網址於控制模組內找尋相對應函式，並進行後端運算，於運算結束後再發送HTTP回應至前端做網頁呈現。

Django現已為一個成熟的網頁開發架構，且具有一定的安全機制，現今許多知名網站也使用Django作為整體開發架構，同時由於其底層為基於python所開發出的網頁架構，能夠很好的與機器學習套件以及模型做串接，故選擇Django作為本次開發的網頁架構。

一張含有 文字, 圖表, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

圖4.2 Django MTV架構基本流程圖

* + 1. 網頁伺服器介紹

為了維持網站運行的穩定性，以及符合目前伺服器上多網站管理的功能，因此將使用Apache2作為網頁管理器，將運行的網站交由Apache2作為代管，藉由Apache 2代管的網站能夠具有更高的運行穩定度，同時也能避免不同網站間的衝突，當網站流量較大時，也能夠進行流量分配，提高網站性能與擴展性，於此Apache2將進行DNS轉址、網頁加密、網址管理等工作。

網頁伺服器除了上述功能外，於安全性上也具有其重要性，藉由網頁伺服器代管的網站同時保護伺服器與用戶間的資料安全，監測以及阻擋來自外部的惡意攻擊，阻擋不當訪問以及保護伺服器不受到網路威脅。

本次實驗將以Ubuntu作為作業系統，並於Ubuntu作業系統上運行Apache 2作為網頁伺服器加以管理網站。

4. 2. KDPredictor網頁介紹
      1. 單病患輸入區域介紹

本次KDPredictor以網頁工具作為呈現，以單頁式方式呈現，網頁上包含有兩種輸入模式的輸入區域，分別為將單病人抽血資訊以於網頁上填入選單的方式作為單病患輸入，以及將單個或多個病患的抽血資訊以.csv檔方式作為輸入的多病患輸入區域，當中多病患輸入區域須按照提供的範例檔案給定欄位名稱，兩種輸入方式個特徵輸入時須注意單位是否一致，若病患無明確看診日期，則可以當天日期作為替代方案輸入。

一張含有 文字, 螢幕擷取畫面, 行, 數字 的圖片

自動產生的描述

圖4.3 單病患輸入區域

於單病患輸入方式，共提供18個特徵欄位提供使用者輸入，於模型所需要的特徵當中的年齡會由出生年月日與看診日期作轉換，於WBC\DC檢驗中的(%)欄位會經由與CBC檢驗中的WBC(1000)欄位做計算後得出每個白血球的真實計數(count)，同時Eosinophil、Hemoglobin、Platelets會以年齡為基準，於後端計算出依照年齡分群的Z分數，其餘特徵會於後端依照訓練集的平均值以及標準差轉換特徵數值為Z分數後送入模型預測，於網頁後端會自動辨別輸入的特徵所能運行的模型，並於結果欄位作呈現。

一張含有 文字, 字型, 螢幕擷取畫面, 圖形 的圖片

自動產生的描述

圖4.4 網頁運作流程

* + 1. 多病患輸入區域介紹

於多病患上傳檔案的輸入方式中，目前KDPrdictor網頁工具僅接受使用者使用.csv檔案格式上傳，使用者需提供與單病患輸入一樣的18個特徵欄位如下，當中欄位開頭名稱需與範例檔案相同，年齡須以「-」作為分隔，並以yyyy-mm-dd方式作為輸入格式，所有欄位輸入時資料內單位需與單病患輸入相同，若資料使用單位不同會導致模型預測發生誤差，無法顯現應有的機率給予參考。

1. Born
2. Consultation
3. WBC(1000)
4. RBC
5. Hemoglobin
6. Hematocrit
7. MCV
8. MCH
9. MCHC
10. Platelets
11. Seg (%)
12. Lym (%)
13. monocyte (%)
14. eosinophil (%)
15. basophil (%)
16. Band (%)
17. ALT/GPT
18. CRP

同單病患輸入相同，於模型所需要的特徵當中的年齡會由出生年月日與看診日期作轉換，於WBC\DC檢驗中的(%)欄位會經由與CBC檢驗中的WBC(1000)欄位做計算後得出每個白血球的真實計數(count)，同時Eosinophil、Hemoglobin、Platelets會以年齡為基準，於後端計算出依照年齡分群的Z分數，後續後端運算流程同圖4.4，多病患輸入檔案中各欄位內允許缺值的存在，網頁工具後端會依照現有特徵決定所能運算的模型。

一張含有 文字, 行, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

圖4.5 多病患輸入區域

* + 1. 模型預測結果輸出區域

圖4.6中展示圖4.3以範例輸入後的結果呈現表，當中由於在圖4.3中ALT欄位為空白，因此在模型應用上僅有使用年齡+CBC、年齡+CBC+WBC\DC以及年齡+CBC+WBC\DC+CRP三種模型能夠使用，因此在圖4.6中僅這三個模型於預測結果以及預測機率欄位中有字樣或數值呈現，於年齡+CBC+WBC\DC+ALT、SHAP前五名縮減模型以及全特徵模型則於這兩個欄位已橫槓做表示，網頁每一頁作為呈現單一病患六種模型的各自預測結果以及機率值，同時會於模型類型欄位後呈現輸入的每個特徵值，並且多加提供藉由WBC\DC檢驗中(%)欄位搭配CBC檢驗中WBC(1000)欄位所計算出的白血球真實計數與依年齡所計算的三個z分數值(Eosinophil, Hemoglobin, Platelets)。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

圖4.6 模型範例輸出結果

* 1. 案例探討

本次實驗中與國內長庚醫療機構內的川崎症專家郭和昌醫師合作，將本次實驗所實作的川崎症預測網頁工具投入實際應用，並於臨床上具有一定成效，能夠於傳統川崎症檢定中的臨床症狀發作前，便僅依靠常規血液檢查預測川崎症的發生機率，並且給予醫師參考資訊，下列呈現兩個於臨床上的案例作為範例。

案例一中病患為一個年齡約兩歲的男童，於發燒第二日時就診，就診時並無明顯川崎症典型症狀，而此時進行包含CBC檢驗、WBC\DC檢驗以及生化檢驗三種血液常規檢驗，並將常規血液檢驗數值輸入至網頁工具，如下圖4.8所示，在該案例內儘管是在實驗流程中與其餘模型效能差距最大、表現最差的僅依靠CBC常規血液與年齡檢查的模型，也能夠提供90%為川崎症的機率資訊給予醫師，而表現最佳的全特徵模型則能夠給予高達99.3%為川崎症的機率資訊給予醫師，其餘模型同樣給予96%以上為川崎症的機率作為資訊提供醫師參考。最終該案例符合典型川崎症診斷手段中的發燒四天、眼睛紅腫、皮膚疹、四肢水腫、肛門脫皮、頸部淋巴結0.5cm，以及額外的卡介苗處紅腫資訊，依照這些症狀診斷為川崎症，該案例很好的展現本次實驗所希望達到的目標以及模型效能，在傳統診斷手段中臨床症狀發生之前，便能夠給予醫師可靠的參考依據，使醫師於後續診斷流程中，能夠有一個較為明確的診斷方向。

一張含有 肉, 人員, 手指, 指甲 的圖片

自動產生的描述

一張含有 文字, 腹部, 螢幕擷取畫面, 人員 的圖片

自動產生的描述

圖4.7 案例一臨床症狀圖

一張含有 文字, 收據, 數字, 字型 的圖片

自動產生的描述

圖4.8 案例一驗血數值與模型預測結果

案例二中為一個年齡約2歲的女童，於典型川崎症臨床症狀上具有紅眼、嘴唇病變、四肢脫皮與發燒等症狀，並伴有卡介苗處紅腫發生，該案例同樣進行包含CBC檢驗、WBC\DC檢驗以及生化檢驗三種血液常規檢驗，並將常規血液檢驗數值輸入至網頁工具，如下圖4.10所示，在該案例中於實驗流程效能最差的僅依靠年齡與CBC建立的模型，給予醫師94.1%為川崎症的機率作為資訊，表現最佳的全特徵模型，則給予醫師97.9%為川崎症的機率作為資訊，其餘模型同樣給予94%以上為川崎症的機率作為資訊提供醫師參考，可顯現模型與網頁工具，能夠有效預測川崎症病患，並提供醫師一個簡易方便使用的介面，做為快速預測的參考輔助工具。

一張含有 皮膚, 人員, 管風琴, 牙齒 的圖片

自動產生的描述

圖4.9 案例二臨床症狀圖

一張含有 文字, 收據, 數字, 字型 的圖片

自動產生的描述

圖4.10 案例二驗血資料與模型預測結果

* 1. Docker工具介紹

由於部分醫療機構座落位置偏遠，未必具有穩定的網路連線功能，或設備老舊導致連線有問題，為了解決該問題，本次實驗將能夠直接運行的網頁程式碼以及環境包裝成Docker，提供使用者能夠直接下載。

Docker為一種虛擬環境平台，相較常見的虛擬機而言，Docker平台能夠在不建立新作業系統核心的情況下，提供與開發環境完全相同的平台，以此來達到輕量化的目的，於Docker內包含建立作業系統以及環境內所需的映像檔(image)，以及藉由映像檔所建立出的容器(container)，使用者可藉由從官方儲存庫中下載環境所需映像檔，並於由該映像檔所立容器運行想要的專案。

一張含有 文字, 螢幕擷取畫面, 字型, 標誌 的圖片

自動產生的描述

圖4.11 Docker基本架構展示圖

在本次實驗內在Docker映像檔內建立Ubuntu系統，並將由Django架構建設的網頁工具運行於該Ubuntu系統內，使用者僅需執行Docker指令於官方儲存庫內下載映像檔，並進入環境內的KD資料夾中以runserver的方式執行Django環境的manage.py檔案，即可以於本地端在不依靠外部網路的方式運行本網頁工具，由此便可以幫助偏遠醫療機構，在診斷資源以及經驗不充足的情況下，給予醫師一個能夠早期提供川崎症發病機率資訊的強力工具。

5. 結論與未來展望



10. 1. 結論

本次研究依照血液抽血樣本處方單據以及特徵重要性分析提供六種不同模型，分別為使用年齡+CBC+WBC\DC檢驗以及生化檢驗的全特徵模型、年齡+CBC檢驗、年齡+CBC+WBC\DC、年齡+CBC+WBC\DC+ALT、年齡+CBC+WBC\DC+CRP以及SHAP分析前五名特徵縮減模型六種模型，於測試集內分別獲得0.886、0.778、0.835、0.847、0.886與0.886作為招回率，並藉由OLS迴歸分析檢驗，驗證本次研究模型與臨床症狀的相互獨立性，此外更藉由來自同機構不同時間段、不同機構、不同國家等獨立驗證資料驗證本次研究模型的泛用性。

除去單純數據分析外，本次研究以Django架構作為基礎，建立一個簡易操作的網頁工具KDPredictor，該網頁工具提供使用者輸入單病患或多病患檔案功能，同時會由後端辨識可使用模型，並依照可使用模型分別於前端呈現六種不同模型中可使用模型的預測結果以及機率值，提供醫師一個快速並簡易使用的網頁工具，用於早期診斷時能有一個參考依據。

5. 2. 未來展望

現今本研究已經提供一個包含依照常規血液檢查處方單據所建立的多種機器學習預測模型原型網站，並且以將其以Docker專案型式包裝，然而仍舊於內部使用與測試階段，希望於將來能夠將該網站不屬於各醫療機構內，或開放使用這自由下載，以此便能夠提供偏鄉低區或對於川崎症並不熟悉的醫師一個有用的參考資訊，進而於早期診斷出川崎症病患，減少後天心臟病的發生可能性。

1. 附錄
2. 參考文獻

[1] <https://www.cmuh.cmu.edu.tw/HealthEdus/Detail?no=5079>

[2] <https://www.mmh.org.tw/child/know_health_view.php?docid=73>

[3] KD biomarker paper

[4] KD CNN與圖片paper

[5] lings

[6] lams

[7] tsai

[8] <https://github.com/numeristical/resources/blob/master/CalibrationWorkshop/Calibration_Workshop_1.ipynb>