HW2\_Report

系級:智能系統 姓名:張宸瑋 學號:312581006

1. Data pre-preprocessing and any other data-centric procedures you conducted

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1. Dataset analysis  |  | | --- | |  | | 訓練資料標籤分佈狀況直方圖 | |  | | 資料缺失值統計 | |  | | 每個特徵對於死亡的Spearman相關系數直方圖 | |  | |  | |  | |  | | 透過Pandas Profiling做探索性數據分析(因為頁面空間，只顯示部分結果 | |
| 1. Feature Selection  |  | | --- | | **方法一: 過濾式特徵選取** | |  | | 對於數值型的資料，透過pearson相關系數獲取與目標特徵為強關聯性的特徵。 | |  | | 對於類別型的資料，透過卡方檢定，獲取與目標特徵之間的關聯性，判斷是否為相互獨立，抑或是有相依的關係存在。 | | **方法二: 嵌入式特徵選取** | |  | | 透過隨機森林能夠為每個特徵提供分數，這些分數衡量了每個特徵對於模型的性能有多大的貢獻，透過這些特徵重要性來進行特徵選取。 | | **方法三: 降維方法** | |  | | 透過主成分分析，將原本的特徵空間轉換為新的特徵空間，保留最重要的訊息。 | | 實驗結果:  在我最後的實驗結果中，透過隨機森林的方式選擇特徵，獲得了最好的表現。 | |
| (3)Data cleaning   |  | | --- | |  | | 透過孤立森林處理異常值 | |  | | 在資料分析階段，我觀察到encounter\_id以及patient\_id的每個值 都是唯一的(圖1)，因此選擇把這些特徵從原始資料刪除。 | |  | | 圖1 | |
| 1. Data transformation  |  | | --- | |  | | 對於數據集中的分類特徵做One-Hot Encoding，並且數值特徵做標準化。 | |
| 1. Data imputation  |  | | --- | |  | | 對於數據集中的缺失值，數值特徵透過線性插值填充，並且針對分類特徵，計算該分類特徵的眾數，也就是以出現頻率最高的值進行填充。 | |
| 1. Data imbalance handling  |  | | --- | |  | | 在資料Imbalance方面我透過Tomek Links來緩解資料不平衡的問題，其原理為如果有兩個不同類別的樣本，他們分別是彼此之間的最近鄰，也就是A的最近鄰是B，B的最近鄰是A，那麼A、B就是Tomek Links，而如果其中一個組成Tomek Links的點屬於多數類的樣本，那就把他刪除，示意圖如下。 | | https://lh7-us.googleusercontent.com/7UM2GjnijoaRq6T473RCgI3vffXWZOfmkz019hnVt9wSuqcjSbirGt3tKKOnYZ6JPJOVGnYSGRFriX8_cOb3ItMcJypFHtqxb-gdjaYZlxviPHCOziJT1hsiTII3O4AlcJKCCSF7Y6QvKbdcTvINSAMsSA=s2048 | | Tomek Links示意圖 | |
| 討論:   1. **discuss problems encountered**   在資料分析時，我發現資料有非常嚴重的imbalance問題，並且有不少的特徵都有缺失值的情況。   1. **explain how you deal with them**   在資料imbalance方面，我嘗試了不同的演算法，像是Tomek Links、Smote+ENN以及Smote+Tomek Links，來緩解資料不平衡的問題，最後的實驗結果顯示Tomek Links的效果最好。  而在資料缺失值方面，對於分類特徵我只採用了透過計算該分類特徵出現頻率最多的值，來填補缺失值，而在數值特徵方面，我嘗試使用眾數、線性插值、以及平均值來填補缺失值，最後實驗下來的結果是線性插值最好。 |

(2) Classification Methods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1. Describe clearly the machine learning algorithms (give References to the used algorithms) and related packages you used.  |  | | --- | | Describe clearly the machine learning algorithms | |  | | 這次我使用的是XGboost演算法，全名為 eXtreme Gradient Boosting，這個演算法是目前最常見的演算法之一，它是以Gradient Boosting為基礎下去實作，並增加一些新的技巧，並且結合了Bagging跟Boosting的優點。  在XGboost中，每一棵樹是互相關聯的，目的是希望可以藉由後面生成的樹能夠修正前面一顆樹犯錯的地方，且採取了隨機採樣的技巧，和隨機森林一樣，在生成每一棵樹的時候，隨機提取特徵，因此每棵樹的生成並不是由所有的特徵參與決策，並且XGboost在目標函數中添加標準化，這樣有助於幫助模型控制複雜度，並且避免過擬合，也提高了模型的泛化能力。  並且XGboost能夠計算每個特徵的重要性分數，幫助我們能夠理解不同特徵對於分類目標特徵的貢獻程度。 | |  | | 這次的實作，我使用到了xgboost庫，並且使用了它所提供的XGBClassifier類別來完成這次的分類任務。 | |
| 1. Describe how to reproduce the results with your source code files.  |  | | --- | |  | | 透過執行main.py生成預測結果 | |  | | 執行完後會生成出test\_pred.csv | |  | | 程式所對應的參數分別是test\_X.csv的位置、輸出test\_pred.csv的位置以及sample\_submission.csv的位置 | |  | | 注意:要將model.json放到跟main.py同一層資料夾下 | |

1. Results & Analysis

|  |  |  |
| --- | --- | --- |
| 1. Count Average AUROC and macro F1-Score with cross-validation method.  |  | | --- | |  | | 結果截圖 | |
| 1. Explainable experiment: |
| * Show the Top 20 most important features. |
| * Give some analysis and insights about the high-importance features.   透過上方特徵重要性的圖表中我們可以看到最重要的兩個特徵為pre\_icu\_los\_days以及icd\_id，上網查詢這兩個特徵代表的意義好像是在病患進入icu前在醫院內的住院天數以及不同的重症監護單元標識符，我認為以這個來作為判斷依據算是有道理的，並且在前幾個重要的特徵中透過bmi以及age這種較為基本的病患資料，也能夠符合我們的想法，因為如果年紀越大，病患死亡的機率確實有可能更高，但我比較意外的是bmi以及age的特徵重要性居然能排那麼前面，因為在它們後面的特徵看起來像是病患的一些更詳細的身體特徵，透過這次的實驗我認為透過像XGBOOST這種具有可解釋的模型，能夠讓我們挖掘到一些我們可能沒辦法觀察到的特徵重要性，並且也可以讓我們了解到模型是如何預測，並且追蹤模型的決策過程。 |

Reference:

1. <https://zhuanlan.zhihu.com/p/142413825>
2. <https://medium.com/@emilykmarsh/xgboost-feature-importance-233ee27c33a4>
3. <https://zh.wikipedia.org/zh-tw/XGBoost>