\*以joineR：：heart.valve資料為主先嘗試

Training 0.7 test 0.3 test new data

**模型方法：**

1. BIMM RF (Binary outcome
2. RF
3. GLMM
4. GEE

**建模過程：**

* 11/25 起始進度：BIMM RF初步嘗試 => 沒辦法收斂
* 11/25

先確認類別變數的排列組合 沒有人數是0的狀況 再run一次

還是沒辦法收斂 變數lvmica~sex+age+time+grad+ef+bsa

* 11/27

看code為甚麼不能收斂 =>原因random effect variables 放入方式錯誤

+( 1｜time）+（1 | age) 隨機效應變數用|區隔

Data : heart.valve

Formula : lvmica~sex+age+time+grad+ef+bsa

Training : train.preds 0acc : 0.699 ; 1acc : 0.942 ; acc : 0.862

0 1

0 158 68

1. 27 439

test.preds 0acc : 0.32 ; 1acc: 0.9 ; acc : 0.71

1 2

0 32 67

1. 18 179

* 11/28

Try H1 H3的model 看他的收斂方式 比較3個的acc sen spe

H1: 迭代2次

train.preds 0acc : 0.796 ; 1acc : 0.965; acc : 0.91

0 1

0 186 46

1 16 449

test.preds 0acc:0 1acc:1 acc:0.66

0 1

0 0 99

1 0 197

* 12 / 7

H3 model迭代一次

train.preds 0acc : 0.75 ; 1acc : 0.95; acc : 0.89

0 1

0 171 55

1 20 446

test.preds 0acc:0.57 1acc:0.42 acc:0.47

0 1

0 57 42

1 113 84

RF train.preds train0acc:0.41 train1acc:0.89 trainacc:0.73

0 1

0 94 132

1 50 416

Testprob test0acc: 0.33 test1acc: 0.91 testacc:0.72

0 1

0 33 66

1 17 180

**下次:**配合paper看符號對應到的code variable是誰

改用 單一資料的new obs做test

隨機效應有放錯 可能要改一下 了解nested cross random effect的差異

可以試試看normalization（變成0-1）效果可能會比較好

試比較GLMM RF的跟原本paper的結果有沒有類似

\*\*想碩論動機跟貢獻

\*\*在甚麼情況會比較好

Y 可能為多類 用兩類分類器的策略: one against one / one against all

* 11/25 起始進度 HBP dataset已初步整理好

路徑：C:\Users\hsu\Desktop\master\TCHCData\hbp\_devtvs.csv

**待作清單：**

1. 決定預測目標(Y)
2. 挑選適合重複測量的插補資料方法 先看HBP欄位NA比例
3. 挑選變數
4. 輸入不同的input 比較模型結果

* 下次：

1. 決定預測目標(Y)

用hbp\_devtvs.csv裡的**檢測值欄位NA的比例**進行挑選

結果:

總筆數裡NA 比例最低的6個 (和血壓沒有關係的) 總筆數3032

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 變數 | HR | Total\_C | Creatinine | TG | LDL | glu\_AC |
| NA比(筆數)/有值筆數: | 44.49%  (1349)  1683 | 57.16%  (1733)  1299 | 57.49 %  (1743)  1289 | 58.91%  (1786)  1246 | 60.32 %  (1829)  1203 | 60.75 %  (1842)  1190 |

|  |  |  |  |
| --- | --- | --- | --- |
| 變數 \NA比例(人數)/有值筆數 | V1 Total 1378 | V2 Total 674 | V3 Total 346 |
| HR | 18.43% (254) 1124 | 67.8% (457) 217 | 17.63% (61) 285 |
| Total\_C | 32.15%(443)935 | 66.02%(445)229 | 64.45% (223) 123 |
| Creatinine | 33.02%(455)923 | 65.73%(443)231 | 63.87%(221)125 |
| TG | 34.03%(469)909 | 69.29$(467)207 | 65.90%(228)118 |
| LDL | 37.01%(510)868 | 68.25%(460)214 | 67.63%(234)112 |
| glu\_AC | 38.32%(528)457 | 67.8%(457)217 | 67.34%(233)113 |

因為模型是預測2類變數=>連續變數切成2類變數

和血壓有關係的NA比例最低的 :



1. 挑選適合重複測量的插補資料方法 先看**HBP欄位NA**比例

結果：

|  |  |  |  |
| --- | --- | --- | --- |
| HBP(總人數) | V1(1507) | V2(832) | V3(487) |
| 完整數值人數 | 368 | 223 | 124 |

7天\*4 =28 收縮 舒張共56筆 +平均6筆

56筆中超過3天沒有數值的(NA >12\*2)

|  |  |  |  |
| --- | --- | --- | --- |
| HBP(總人數) | V1(1507) | V2(832) | V3(487) |
| 超過3天沒有數值的 | 155 | 32 | 17 |

插補方法: k-NN 隨機森林 平均 中位數 …

1. 其他變數挑選

可能依照NA比例、回歸裡的顯著性…

1. Data分訓練測試的部分，因為是重複測量資料

測試是new patient

測試是訓練資料的new observation

1. TCHC 7天HBP (引用來源)

7天取6天 預測

7天早上 預測

**12/9與鄭醫師討論結果：**

1. Y用abpm裡的指標(morning surge 早晨血壓竄升 / nocturnal dipping 夜間血壓不降)
2. 白天和夜晚的時間採Long clock-time period的假設：

Daytime : 06:00-22:00 Night-time : 22:00-06:00

1. morning surge 有4種計算方式 目前沒有固定哪一種 好像算出來之後沒有一個判斷的臨界點=>要問 =>台灣沒有標準
2. nocturnal dipping 算出來後有4個階段(Reverse dipper / Non‐dipper / Dipper/ Extreme dipper)
3. if Y = morning surge預測 有沒有 “晨峰血壓增高”=>因為沒有一個切點=>不預測

if Y = nocturnal dipping 預測 Reverse dipper / Non‐dipper / Dipper/ Extreme dipper 4類中哪一類

* 12/12

待做：Data : ABPM

1. 計算TCHC 的 nocturnal dipping rate

3. 和hbp\_devsvt合併

* 12/15跟老師咪挺：把資料弄出來一切好說

先用hbp abpm 合併一起試試看 （其他covariate之後再補就可以）

把nocturnal dipping rate算出來（有問題的四百多個case先跳過）

先整理出complete case子集 試跑 4個模型 gee glmm bimmrf rf

nocturnal dipping 四類用兩類分類器預測 看case分佈決定 都可以嘗試看看

看完結果之後再看遺失值要怎麼插補

插補的策略 1都不補 2補ｘ（hbp） 3補y(abpm) 4兩個都補

去比較四種subset 預測的結果 如果都差不多表示遺失值沒有影響研究 有一致性

插補的同時就差不多要挑變數

插補的方法 再看看

框架類似銘洋學長的做法 可以看他怎麼做一堆模型比較的

* 12/22

目前完成

1. 計算TCHC 的 nocturnal dipping rate 還有status(abpm\_Y.csv) 1670\*194
2. 和hbp合併

hbp\_dip.csv (以Y為主合併) 1670\*125

hbp\_dip\_byx.csv(以X為主合併) 3032\*125

1. 計算hbp\_dip\_byx.csv 中status在各種情況下NA情形

|  |  |  |  |
| --- | --- | --- | --- |
| Extreme dipper | 12 | 198 6.5% | 845 |
| Dipper | 186 |
| Non dipper | 439 | 647 21% |
| Reverse dipper | 208 |
| NA | 2187 | 2187 72% |  |
| Total |  | 3032 |  |

各個訪視內的情況

**V1**

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 9 | 129 8.5% |
| Dipper | 120 |
| Non dipper | 301 | 447 29.6% |
| Reverse dipper | 146 |
| NA | 931 | 931 61.7% |
| Total |  | 1507 |

**V2**

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 2 | 36 4.3% |
| Dipper | 34 |
| Non dipper | 72 | 108 12.9% |
| Reverse dipper | 35 |
| NA | 689 | 689 82.8% |
| Total |  | 832 |

**V3**

|  |  |  |
| --- | --- | --- |
| Extreme dipper |  | 24 4.9% |
| Dipper | 24 |
| Non dipper | 43 | 61 12.5% |
| Reverse dipper | 18 |
| NA | 402 | 402 82.2% |
| Total |  | 487 |

=>有Y的樣本數少 兩類樣本數差距大

Y 是NA =>用分群方式 或是樹的方式產生(會dependent on X) 或是插補ABPM 再算Y

2類樣本數差距大 => resampling (把少的類別放大到跟另一類一樣多)

目前資料(HBP + dipping status)完整樣本數：245

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 3 | 52 |
| Dipper | 49 |
| Non dipper | 143 | 193 |
| Reverse dipper | 50 |

* 12/22 咪挺結果

從 845筆有Y的去做 =>收縮 舒張 早上 晚上 平均 =>取4個(用平均可以解決時間性的問題)

先抓平均完的complete case 去做模型(完整case會變多) 看效果

再對X先做隨機森林插補 再抓4個平均 再跑模型 看效果

做完再去做 對ABPM插補 再算Y 再跑模型

之後再考慮要不要加其他變數

* 12/26

hbp\_dip\_byx 有Y的845筆

Complete case 830筆 564 人

* 1/5咪挺結果
* 1/12

要預測的目標應為Non dipper Reverse dipper (1)VS Extreme dipper dipper (0)

或是Reverse dipper (1) VS其他(0)

* 
* 1/18和鄭醫師咪挺結果 (0118醫師討論ppt)

1. 2vs1把non-Dipper拿掉不對 關心的是non-dipper和reverse dipper
2. 可以加的變數=>血壓變異程度 sd arv dv / heart rate / hba1c / 年齡性別/有沒有投某藥
3. 模型可以用CNN之類的
4. 可參考另一篇paper 裡的變數(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8639874/>)

* 1/20 把原本的TCHC進度 和model進度合併了

預計完成 : 改寫code (未完成)

包括 模型的function 資料切割的過程

* 1/24

預計完成 : 改寫code

包括 模型的function 資料切割的過程

* 1/25進度目標:

1. 模型可以用迭代 (完成)
2. 改用V12 預測V3 (完成)
3. 同個時間點的分析(預測new cases)
4. **把random effect的放法相關網頁整理一下 放法的解釋和差異**
5. **解決test 準確度低的問題**
6. 加入其他變數 => 整理變數 試分析
7. **查mixed回歸裡的加權方法(解決不平衡資料的問題)**

Ex: 在模型裡讓資料平衡 EX:logistic weighted regression

weight mixed model 讓有病沒病的人加權 讓權重balance一點

* 1/26 咪挺內容 / 結果

鄭醫師認為2vs1 (把non dipper刪掉的情況不合理)

更改random effect的放法之後模型表現有好一點

#V1V2 略優於 V12V3

#(1|MRN) 比(1+time|sys)+(1+time|dia)好

#H3H2結果和1iter相同 H1因資料已經偏向1 不需要用

/

把樣本限制在一樣的 看誰表現得比較好 就用那個訓練跟測試資料集

測試sys dia time 22交互作用

如果有交互作用 在考慮要不要放random effect

如果交互作用都不顯著 就放3個變數的random effect

**放鄭醫師提到的covariate**

+(1|MRN)是對的

最後重要的是mixed model 的系數

把係數寫出來 確保自己了解

題目machine learning model for nocturnal dipping (如果最後比較出來的結果是其他模型比較好 就改成optical …..)

* 2/10進度

目標 完成上次咪挺討論出的後續

1. 把樣本限制在一樣的 看誰表現得比較好 就用那個訓練跟測試資料集random = "+(1|MRN)"

V1 V2 VS V12 V3 都限制在一樣的樣本(都是81)

**=> V1 V2表現略優於V12V3 (在V12V3的時候把Visit也加進去?)**

* 2/12 進度

1. **測試sys dia time 22交互作用(data 有標準化)**

如果有交互作用 在考慮要不要放random effect

如果交互作用都不顯著 就放3個變數的random effect

**加了COV COV也要測交互作用項?**

|  |  |  |
| --- | --- | --- |
| GLMM | V12 V3 | V1 V2 |
| Sys \* time | 0.86 | 0.78 |
| Dia \* time | 0.93 | 0.87 |
| Sys \* dia | 0.13 | 顯著 p-value = 0.0494 |

* 2/13

1. 整理covariate 放進去測試

可以加的變數=>血壓變異程度 sd arv cv / heart rate / hba1c / 年齡性別/有沒有投某藥

* 2/18

整裡好的變數=> heart rate / hba1c / 年齡/性別/有沒有投某藥

有沒有時間性的變數 要分開merge 性別年齡才會每一筆都有

**如果年齡性別有NA =>預設刪掉**

* 2/19

插補方法=>隨機森林 (有MissForest ~~和 rfImpute 2種~~)

插補變數=> heart rate / hba1c/CCB

測試插補heart rate / hba1c之後的**結果**

**=>加了其他變數後 H1 H2迭代1000次 =>做標準化可能會比較快收斂?**

**H3迭代100 次**

~~如果要比較插補方法=> 用有數據的資料產生na 插補後看正確率~~

1. 最後重要的是mixed model 的系數 把係數寫出來 確保自己了解
2. **調整 random effect 放法的解釋和差異**
3. **解決test 準確度低的問題 => 可能是測試只用隨機森林的關係?**
4. 查mixed回歸裡的加權方法(解決不平衡資料的問題)

Ex: 在模型裡讓資料平衡 EX : logistic weighted regression

weight mixed model 讓有病沒病的人加權 讓權重balance一點

weighted的方法 inverse / extreme /其他=>查一下

有weighted random forest

1. **加畫ROC curve AUC / 其他COV的交互作用 / resampling**
2. V1V2 V12V3 訓練測試的選擇和解釋

* 2/21

做完加變數後的測試，做PPT

V1V2無法收斂 嘗試做標準化之後再測試一次

標準化後 可以收斂 但效果沒什麼改變

* + 從共線性開始測試 看資料需不需要做調整
* 2/22 PPT

做35-40張左右 重複測量機器學習模型 希望可以提供模型的建議讓臨床使用

從資料介紹 題目動機 文獻回顧 目的

資料前處理 變數介紹 怎麼從3000多筆到800多筆

資料分析方法 隨機效應調整 交互作用 模型數學式 結果

加入變數 插補方法介紹(根據某paper得出 隨機森林插補方法比較好 所以直接用) 結果

結果討論和未來工作

人數 / 筆數 表格

|  |  |  |
| --- | --- | --- |
| 訓練 / 測試 | 沒有covariate | 有covariate |
| 總筆數 | 817筆 | 761筆 |
| V12 / V3 | 324 筆 / 162 筆  81人 | 284 筆 / 142 筆  71人 |
| V1 / V2 (與V12V3人數相同) | 162 筆 / 162 筆  81人 | 142 筆 / 142 筆  71人 |

* 2/24 咪挺內容

1. **把樣本限制在一樣的 看誰表現得比較好 就用那個訓練跟測試資料集**random = "+(1|MRN)"

V1 V2 VS V12 V3 都限制在一樣的樣本(都是81 / 加入變數71)

=> V1 V2表現略優於V12V3 (在V12V3的時候把Visit也加進去?)

1. **測試sys dia time 22交互作用(data 有標準化)**

如果有交互作用 在考慮要不要放random effect

如果交互作用都不顯著 就放3個變數的random effect

**加了COV COV也要測交互作用項?**

|  |  |  |
| --- | --- | --- |
| GLMM | V12 V3 | V1 V2 |
| Sys \* time | 0.86 | 0.78 |
| Dia \* time | 0.93 | 0.87 |
| Sys \* dia | 0.13 | 顯著 p-value = 0.0494 |

1. **整理covariate 放進去測試**

整裡好的變數=> heart rate / hba1c / 年齡/性別/有沒有投某藥

如果年齡性別有NA =>預設刪掉

|  |  |  |
| --- | --- | --- |
| 訓練 / 測試 | 沒有covariate | 有covariate |
| 總筆數 | 817筆 | 761筆 |
| V12 / V3 | 324 筆 / 162 筆  81人 | 284 筆 / 142 筆  71人 |
| V1 / V2 (與V12V3人數相同) | 162 筆 / 162 筆  81人 | 142 筆 / 142 筆  71人 |

插補方法=>隨機森林 (有**MissForest** 和 rfImpute 2種)

插補變數=> heart rate / hba1c / CCB

測試插補heart rate / hba1c之後的結果

=>加了其他變數後 H1 H2迭代1000次 H3迭代100 次

V1V2無法收斂 嘗試做標準化之後再測試一次=> 標準化後 可以收斂 但效果沒什麼改變

* 2/24咪挺結果

把sys dia 換成Sbp / MAP 做一次 在V12V3的時候把Visit也加進去?

試著解釋 V1V2 > V12V3 的原因

為甚麼不做 new patient

* 3/7

1. 檢定有沒有加協變量的樣本下降是否隨機(卡方齊一性檢定)

=> done 樣本下降，有沒有+COV的2者(1/0)比例仍相同

1. 更換seed 調整正確率 =>目前測起來沒有改變(訓練準確率100%的部分 測試有變)
2. 列出隨機森林和隨機森林插補中的參數 => done目前一棵樹只有3個變數 可以再調

* 3/13

1. **對醫院做校正，做法: 加入該變數 => done**
2. **改用lme測試 => 跑完全部補準確度到excel上**
3. 列出隨機森林和隨機森林插補中的參數 =>目前一棵樹只有3個變數 可以再調
4. BIMM 的模型數學式 估計出來的參數 最重要的time變數
5. **把sys dia 換成Sbp / MAP 做一次=>變數做好了 還沒測試 3/14:不用做了**
6. **在建模之前做重抽樣 / 標準化 / 刪掉離群直 調整正確率**
7. 更改隨機效應 調整正確率
8. 針對隨機森林impute做多次合併 之後做挑選變數
9. 畫一個分析流程圖像masked HTN那篇
10. 了解GLMER control的差別 還有調整optimizer: Nelder\_Mead
11. 鄭 : 沒有covariate的分析，單純使用home BP 血壓紀的資訊，可以考慮將不同的Visit 全部混合在一起 使用cross-validation 的方式來訓練及驗證模型

Specificity 太低的情況，看看是否有模型可以改善

1. 被問到的問題 為甚麼不能有遺失值 訓練正確率是1 這些問題還有再BIMM多補 先假設獨立再解決掉相依性的解釋 PPT就留著之後繼續用
2. **口試** : 一個人1小時(報告20分鐘) 寄電子檔(問需不需要紙本檔案) 鄭醫師 程毅豪老師 楊欣洲老師 盡量早上 地點:教學二館4樓 食物 鄭醫師的信分開寄 說只需要勾一個時段(我的) 5/23(一) 5/23(五) 鄭醫師只有下午可以 **禮拜三問郭之後再看看要怎麼約所有老師**

* 3/14 鄭醫師咪挺結果

真的有空再說 : 可以用GYMNS data做external validation =>拿一個時間點就好 找時間來榮總抓資料 =>3/16郭 : 時間不夠 不用

重複測量臨床沒有意義，可以做預測新病人

最好3個指標(準確度 敏感度 特異度)都至少0.6 0.7左右

變數=>改放有沒有糖尿病 (demographic)

和用了幾種高血壓相關藥物(綜合全部的藥物去看他用了幾種)

TCHCData有清得更完整的HBP和vital signs可以用 => 在edited TCHC裡面

=>資料要重新整理 不做

* 3/19

撈出新變數的資料

1. 有沒有糖尿病 (demographic) =>完成
2. 用了幾種高血壓相關藥物(綜合全部的藥物去看他用了幾種)=>完成

Complete data 再做一次 比較有沒有imputation有沒有一致性

撈出V1的資料 一樣分成有COV 沒有COV=>完成

分割所有訓練和測試資料集 在train test file check.r上

所有scenario 都去跑 RF BIMMRF 調整參數 到最漂亮的結果 打在excel上

做oversampling 調整class weight

* 3/26

在try other data.R裡 整理了其他變數+CCTr CCTe上 1048筆

排除掉超過30%遺失值的變數 =>存成COV less30

用complete case 368筆 訓練隨機森林 找變數重要性

* 3/28

用變數重要性 前10個重要的變數(或排序一下找斷點) +上原本的變數 =>model performance島裡面

Case測隨機森林看看 => complete case下，在切了訓練測試之後 OOB error 6-8%左右 ，沒切OOB error大約3-5%

基本上只要隨機森林OK 測試就不會太差了

在隨機森林可以調整class weight wj=n/(k\*nj)

BIMM調thershold 以1或0的比例當作thershold

用caret套件的 random / grid search找mtry的參數

**把加上挑選出來變數的資料集處理好，做完imputation後可以分割**

**Noncov的基本上不用改 可以直接跑model測試**

* 3/30

題目要改成

**The Optimal Machine Learning model to identify nocturnal dipping status ?**