# 肝癌預測

統計學習與資料探勘期末書面
高念慈、林良朋

指導
鍾思齊 教授

December 30, 2022

**Abstract**

近年來受到疫情影響，原本就缺乏的醫療資源，雪上加霜，像醫生這種高專業的職業，面對如此多的病人也只能束手無策，所以要是能夠正確的預測肝癌或是其他跟醫療相關的資料就有機會能讓資源用在眞正需要的人身上，故此爲我們選擇處理這份資料的動機，而過程主要分爲兩大部分，一爲探索和整理資料，二爲建模，嘗試去發現變數跟我們目標之間的關係。

## 1 資料介紹

資料來自 : https://www.cbioportal.org/study/clinicalData?id=lihc$_t cga$

爲一份跟肝細胞癌有關的資料，其中有 379 筆資料，104 個欄位，是收集了不同時期，歷年來有著跟肝癌相關的資料，其中有兩筆資料爲兩位病人，二次檢驗的結果，不過在此我們還是把資料視爲 379 個獨立樣本，104 個欄位也不是每個病人都做了如此多的檢驗，更多的是重複、缺失值，在變數解釋的地方，主要只對最後模型所使用的變數加目標變數做詳細解釋。

### 1.1 變數解釋

目標變數和 25 個主要變數再多加幾個好解釋的變數。

(61) Overall.Survival.Status : (0,247),(1,132), 0表示生存，1表示死亡

- 25個主要變數

(4) Ablation.embolization.tx.adjuvant : 消融(剝蝕).栓塞.tx.佐(輔助)劑，(類別，Yes/No)，消融和栓塞治療是破壞腫瘤的不同方法，用於幫助預防或緩解症狀，並且通常與其他類型的治療一起使用
(6) AFP.Normal.Range.Lower.Bound : 正常範圍下界(int)，
AFP，1.31–7.89 ng/ml（男性）和1.01–7.10 ng/ml（女性）
AFT : 用作肝細胞癌(HCC) 篩查、診斷和治療隨訪的腫瘤標誌物
(10) Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.Code : 腫瘤。疾病淋巴。節點。階段。美國癌症聯合委員會，(類別，NX，N0，N1)
N category 描述癌症是否已經到達附近的淋巴結
(14) Bilirubin.Total : 總膽紅素，正常總膽紅素大約在1.2mg/dL以下，(num.)
總膽紅素: 直接型膽紅素＋間接型膽紅素
(16) Bilirubin.Total.Norm.Range.Upper : 膽紅素.總計.正常。範圍.上限，(num.)

膽紅素數值的重要性，不下於肝發炎指數
(19) Child.pugh.classification.grade : 肝硬化嚴重程度的Child-Pugh 分級，(類別，A，B，C)
數字越小狀況越好，A : 5-6分B : 7-9分C : 10-15分
(30) Specimen.collection.method.name : 標本.採集.方法.名稱，做了怎麼樣的手術
肺葉切除術、肺段切除術(單) 等，(類別，6種)
(32) Disease.Free.Status : 無病狀態，(類別，有無復發)
(38) Family.History.of.Cancer : 癌症家族史，(類別，YES/NO)
(40) Fraction.Genome.Altered : 分數.基因組.改變，(num.)，用於查找與特定改變或臨床屬性（例如生存期或腫瘤分期）的相關性
(41) Neoplasm.Histologic.Grade : 腫瘤.組織學.分級，(類別)
分化良好(由細胞接近正常)至未分化
(43) Adjacent.hepatic.tissue.inflammation.extent.type : 鄰近肝組織炎症範圍類型，
(類別，嚴重，中，無)
(45) History.hepato.carcinoma.risk.factor : 歷史.肝癌.風險.因素，(類別，19 類)
飲酒—乙型肝炎—丙型肝炎、非酒精性脂肪肝等
(56) Mutation.Count : (373,val),(6,NA) 其值爲表示病人體內在不同基因組之間產生突變的總和，稱爲突變計數，可作爲評判腫瘤突變的潛在指標
(60) Overall.Survival..Months. : (378,val),(1,NA), 從診斷開始（或治療開始）到結束觀察的時間(以月爲單位)
(68) Adjuvant.Postoperative.Pharmaceutical.Therapy.Administered.Indicator : (228,NO),
(15,YES),(136,NA), 術後佐藥物治療指標標，(類別，YES/NO)
(69) Platelet.count.preresection : (312,val),(67,NA), 在診斷腫瘤後，切除前，血小板的數值，(int)
(70) Laboratory.prcoedure.platelet.result.lower.limit.of.normal.value : (311,val),(68,NA),
實驗程序中，血小板正常值的下界，(int)
(71) Laboratory.prcoedure.platelet.result.upper.limit.of.normal.value : (311,val),(68,NA),
實驗程序中，血小板正常值的上界，(int)
(74) Laboratory.procedure.prothrombin.time.result.value : (302,val),(77,NA), 凝血活時間測試，(num.)，可檢查您的血漿是否存在凝血因子異常
(76) Laboratory.procedure.international.normalization.ratio.result.lower.limit.of.normal.value :
(273,val),(106,NA), 凝血活時間測試中，(num.)，以INR（國際標準化比率）正常值下界
(87) Laboratory.procedure.albumin.result.specified.value : (305,val),(74,NA), 白蛋白測定結果值
(98) Tissue.Source.Site : (379,val), 組織來源地，每組織對應一種實驗項目，(類別，36種)
(100) Person.Neoplasm.Status : (236,TUMOR FREE),(115,WITH TUMOR),(28,NA),
個人腫瘤狀況，(類別)TUMOR FREE : 腫瘤清除/ WITH TUMOR : 有腫瘤
(103) Patient.Weight : (351,val),(28,NA), 患者體重

- 9個主要變數

(5) AFP.At.Procurement : 採購AFP(甲型胎兒蛋白)，(int.)
(26) Laboratory.procedure.creatinine.result.upper.limit.of.normal.value : 實驗室.程序.肌酸酐.結果.上限.的。普通的。價值，(num.)
增加於：腎絲球腎炎，腎盂腎炎，尿毒症，末端肥大症．巨人症，甲狀腺功能亢進
減少於：衰弱、肌萎縮(因年紀太大或肌肉量少，尤其懷孕前六個月)
(45) History.hepato.carcinoma.risk.factor : 歷史.肝癌.風險.因素，(類別)
(56) Mutation.Count : (373,val),(6,NA) 其值爲表示病人體內在不同基因組之間產生突變的總和，稱爲突變計數，可作爲評判腫瘤突變的潛在指標

(60) Overall.Survival..Months. : (378,val),(1,NA), 從診斷開始（或治療開始）到結束觀察的時間(以月爲單位)

(74) Laboratory.procedure.prothrombin.time.result.value : (302,val),(77,NA), 凝血活時間測試，可檢查您的血漿是否存在凝血因子異常

(98) Tissue.Source.Site : (379,val), 組織來源地，每組織對應一種實驗項目，(類別，36種)

(99) TMB..nonsynonymous. : (372,val),(6,NA), 腫瘤突變負荷(TMB)，定義爲腫瘤基因組每個編碼區的非同義突變總數

(100) Person.Neoplasm.Status : (236,TUMOR FREE),(115,WITH TUMOR),(28,NA), 個人腫瘤狀況，(類別)TUMOR FREE : 腫瘤清除/ WITH TUMOR : 有腫瘤

# 2 相關工作

## 2.1 資料分割

把 379 筆資料拆分爲 8 : 2 ，訓練集 303 位，和測試集 76 位，設 random state = 5。

## 2.2 EDA和資料前處理

### 2.2.1 目標類別不平衡

由(圖 1)可以看出此資料集有目標類別不平衡的問題，爲了避免模型出現預測偏向某一結果的情形，加上我們比較好奇影響死亡的資訊，故在此選擇使用隨機過採樣(Over sampling)[3]，把訓練集資料從存活 191 位比死亡 112 位，共 303 位，修正爲存活 191 位比死亡 191 位，共 382 位(圖 2)。

### 2.2.2 第一輪刪變數

(圖 3)跟(圖 4)則是讓我們了解缺失值的分布和比例，可以看到有些變數甚至全爲 NA ，所以在處理缺失值前，我們先做一輪初步刪變數的動作，主要採取三種判斷標準，第一種爲重複、變異性小的資料，以 382 的 9 成爲標準，像，Cancer.Type.Detailed、Neoadjuvant.Therapy.Type.Administered.Prior.To.Resection.Text、Prior.Cancer.Diagnosis.Occurence 等；第二種爲全 NA 的，像，Specimen.Current.Weight、Specimen.Freezing.Means、Specimen.Second.Longest.Dimension 等；第三種爲有一半以上的缺失值，加上資訊不足，爲了避免亂補值會影響模型故刪除，Cancer.diagnosis.first.degree.relative.number，(一個)，在第一輪一共刪了 53 個變數，資料剩 382 rows× 50 columns 。

### 2.2.3 資料分布

連續變數剩下 27 個，類別變數剩 23 個，在此各挑幾個後面篩選出來的重要變數來觀察。(圖 5、 6)

連續 ， KDE Plot : 由左至右，60.Overall.Survival..Months、87.Laboratory.procedure.albumin.result.specified.value、76.Laboratory.procedure.international.normalization.ratio.result.lower.limit.of.normal.value，76.第一眼感覺會是一個很好分辨、很重要的變數，但觀察後面隨機變數重要性的圖後(圖 10)會發現他比60.還要不重要很多，60.在這裡是重要性最高的變數，87.跟76.的重要性是差不多的，主要還是優先看兩種狀況分布能不能找到左右的區分點，再研究同數值但數量有明顯差距的。

類別 ，長條圖 : 由左至右，32.Disease.Free.Status、45.History.hepato.carcinoma.risk.factor、9.American.Joint.Committee.on.Cancer.Metastasis.Stage.Code、10.Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.Code，這四個變數除了 9.沒被選進最後資料集外，其他變數也是都很重要，而第9.明明跟10.長的類似，重要性還比 10.高卻沒被選上的原因爲在後面我們是採取 backward 的方式挑選變數，挑到 10.時，9.因爲太相似了就被略過了。

最後(圖 7)、(圖 8)、(圖 9)則是利用相關係數(線性)跟 TSNE (非線性)[8]來觀察目標變數跟變數或變數間有沒有甚麼特別的關係，可以看到在目標變數跟變數之間相關係數看不出什麼關係；再利

用 TSNE 觀察變數間的聚類情形時，perplexity = 5 也還是分不太出變數間的相似性，推測可能是體檢這類型的特徵本來就很相似。

#### 2.2.4 Encoding

23個類別變數中，主要分成有次序、兩類跟多類別的特徵，我們對前兩種利用mapping[5]，手動給他們對應的 label ，像上面提到的32.跟10.；而像45.這種類別超過10種又沒有次序問題的變數，我們使用 M-estimate Encoder [? ]M-estimateTargetEncoding，他是 target encoding 的一種，主要是看中他能調 m ，一個影響權重的東西，($weight = \frac{n}{n+m}$)，越大會傾向讓整體平均較有影響力，我們在這裡取 m = 1 ，以45.爲例，因爲我們不想讓讓些少數只出現1或2次，影響肝癌的因子，變得跟那些佔了 2、30% 的疾病因子長的類似，利用這個方法還有一個優點爲，他在看到新資料時照樣能 encoding ，$encoding = weight \times$ 類別平均 $+ (1 - weight) \times$ 整體平均，也是因爲如此，他遇到 NA 會把他視爲同類下去 encoding。

#### 2.2.5 缺失值處理

到此，所有類別變數都變成數值型態，只剩下 NA 還未處理， NA 值的部分，因爲是體檢資料，我們推測缺失值純粹爲無做單項試驗或是人爲漏填居多，所以除了上述 M-estimate Encoder 補的四個變數，剩下的變數都直接採取 pandas 內建的 the nearest 插值法[7][4]。

### 2.3 挑選變數

資料剩 382 rows× 50 columns ，雖然已經少了不少變數，但考慮到 row 跟 column 的比例其實還蠻接近的，在此我們利用隨機森林加 backward selection [1]，挑選變數，作爲後續模型訓練和比較的基礎，最後結果爲第二種 AUC 更好，觀察其中變數也發現較直覺、好解釋，故以此 25 個變數做爲最後資料集。

隨機森林參數，(圖 10)爲變數重要性 :

(1) max_depth = 5
(2) max_features = 'sqrt'
(3) max_samples = None
(4) criterion = 'gini'
(5) oob_score = True
(6) bootstrap = True

第一種 : 先挑出前 40 個重要變數，減少重要變數還沒選到就中途停止的情況，
再利用 backward selection 看 AUC 決定最好的變數組合。
結果 : 變數剩 9 個 AUC 分數 : 0.9287626962142198
第二種 : 直接 50 個變數下去 backward selection 看 AUC 決定最好的變數組合。
結果 : 變數 25 個 AUC 分數 : 0.9303714752468215

## 3 模型

### 3.1 10-Fold CV

利用 10-Fold CV 且採用 DecisionTreeClassifier 針對不同的樹的深度及最大葉片個數找出在哪個深度底下ROC_AUC做的最好

最終我們使用 max_depth 爲 4 與 max_leaf_nodes 爲 8 爲後續模型超參數的選取

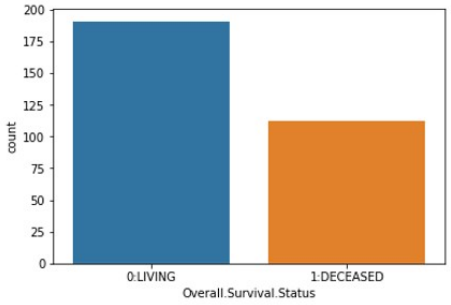


Figure 1: 191 v.s. 112

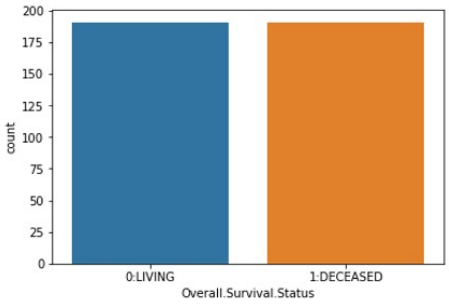


Figure 2: 191 v.s. 191

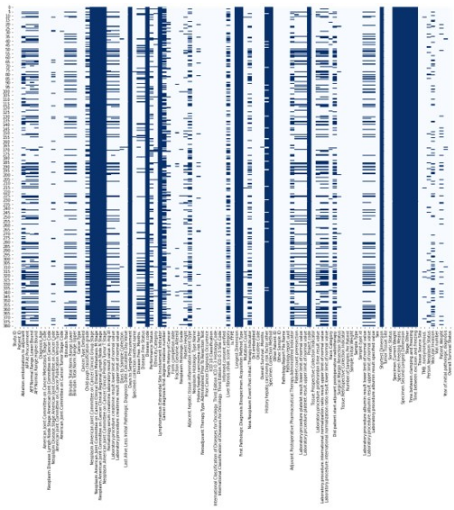

Figure 3: 缺失值分布

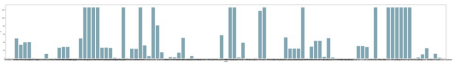

Figure 4: 缺失值比例

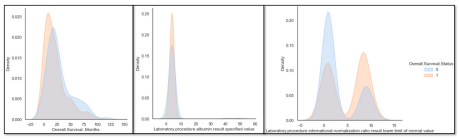

Figure 5: KDE plot

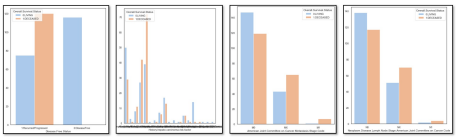

Figure 6: 長條圖

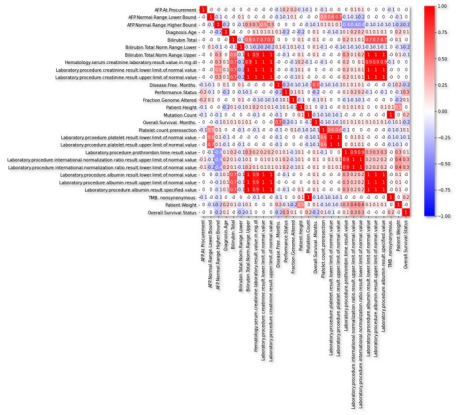

Figure 7: 連續變數相關係數

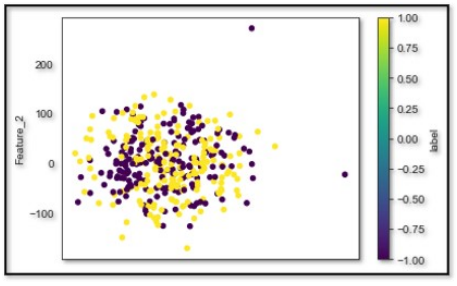


Figure 8: TSNE，二維，50變數

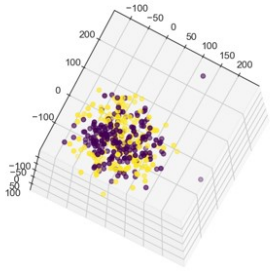

Figure 9: TSNE，三維，50變數

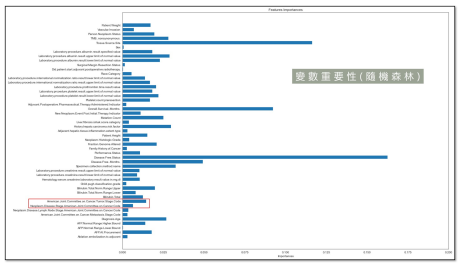


Figure 10: 隨機森林變數重要性

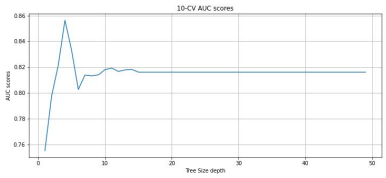


Figure 11: 10 FOLD CV in Tree depth

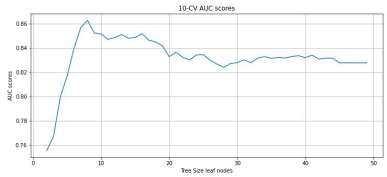


Figure 12: 10 FOLD CV in Tree nodes

## 3.2 Lightgbm

### 3.2.1 Hyperparameter tuning

LightGBM參數選擇(原始投影片得到的結果) [6]

(1) Learning_rate : 0.1 設置越低，對訓練集的學習越深，花費時間越久
(2) max_depth : 4 經由前面所使用的 10-Fold CV 找到的樹的深度，設定太大可能會導致過擬合
(3) num_leaves : 30 沒有使用 10-Fold CV，去找最大葉片個數，設定太大可能會導致過擬合
(4) random_state : 5
(5) n_estimators : 86 以 LightGBM CV 去找出最好的 n_estimators 爲 56

LightGBM參數選擇

(1) Learning_rate : 0.1 設置越低，對訓練集的學習越深，花費時間越久
(2) max_depth : 4 經由前面所使用的 10-Fold CV 找到的樹的深度，設定太大可能會導致過擬合
(3) n_estimators : 10 boosted trees的個數，本身資料集不大，因此使用較小的樹木總數
(4) num_leaves : 8 同樣使用的 10-Fold CV，去找最大葉片個數，設定太大可能會導致過擬合
(5) random_state : 5

### 3.2.2 Feature importance plot

### 3.2.3 Confusion matrix plot in LightGBM

## 3.3 Random Forest

### 3.3.1 Hyperparameter tuning

Random Forest參數選擇
(1) max_depth : 4 經由前面所使用的 10-Fold CV 找到的樹的深度，設定太大可能會導致過擬合

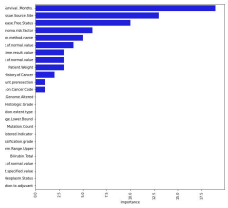

Figure 13: Feature importance plot in LightGBM

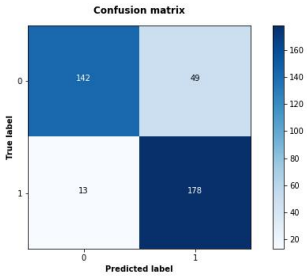


Figure 14: Training set plot

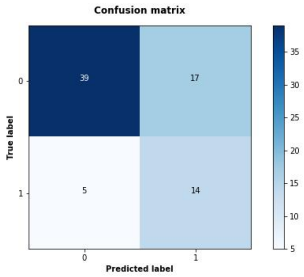


Figure 15: Testing set plot

(2) n_estimators : 10 森林中樹木的數量，本身資料集不大，使用較小的總數(原投影片使用 100)
(3) random_state : 5
(4) max_features : 'sqrt' 考慮特徵數量(平方根)，避免每次都選用所有特徵(原投影片選用所有特徵)
(5) min_samples_leaf : 20 葉節點所需的最小樣本數，也就是至少有 20 個樣本才會做爲終端節點，可以防止過擬合(更新模型多加的參數)

### 3.3.2 Confusion matrix plot in Random Forest

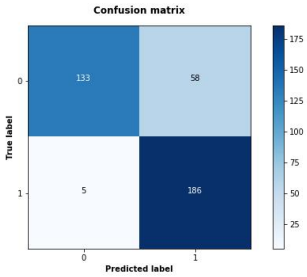


Figure 16: Training set plot

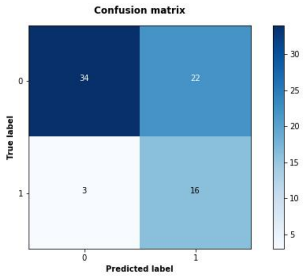


Figure 17: Testing set plot

## 3.4 CatBoost

### 3.4.1 Hyperparameter tuning

CatBoost參數選擇 [2]
(1) Learning_rate : 0.1 設置越低，對訓練集的學習越深，花費時間越久
(2) depth : 4 經由前面所使用的 10-Fold CV 找到的樹的深度，設定太大可能會導致過擬合
(3) n_estimators : 10 boosted trees的個數，本身資料集不大，因此使用較小的樹木總數(原投影片結果爲使用 GridsearchCV 找到最好的顆數爲 86)

(4) l2_leaf_reg : 3 $l2$ 正則化係數選用default值爲 3
(5) random_state : 5

### 3.4.2 Feature importance plot

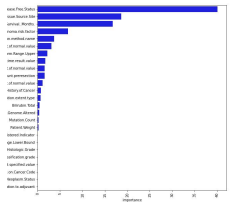

Figure 18: Feature importance plot in CatBoost

### 3.4.3 Confusion matrix plot in CatBoost

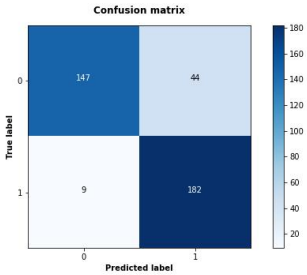


Figure 19: Training set plot

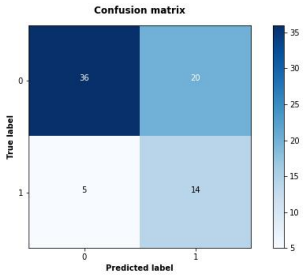


Figure 20: Testing set plot

## 3.5 Ensemble

將LightGBM, Random Forest, Catboost取出其預測機率值平均

### 3.5.1 Confusion matrix plot in Ensemble

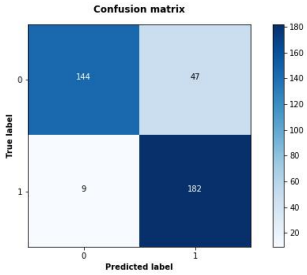


Figure 21: Training set plot

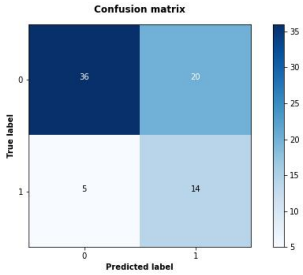


Figure 22: Testing set plot

### 3.6 解釋與討論

- Feature importance : 根據 Figure10, Figure13, Figure18 可以知道在 LightGBM 中，前三大特徵重要性爲 Overall.Survival..Months., Tissue.Source.Site, Disease.Free.Status ; 另一方面，在 Random Forest 與 CatBoost 中，前三大特徵重要性皆爲 Disease.Free.Status, Tissue.Source.Site, Overall.Survival..Months.，顯示這三個變數對於模型的預測結果影響最大

- Confusion matrix : 我們將Threshold調整爲0.4，主要原因是爲了要讓我們實際情形爲死亡我們的預測卻是活著，而導致假陰性(False Negative)的問題，爲了使得其比率下降，進而調降Threshold

## 4 結果

### 4.1 AUC score (以新模型結果爲表格呈現)

| 模型比較 | Logistic | LightGBM | Random Forest | CatBoost | Ensemble |
|---|---|---|---|---|---|
| Training | 0.686001 | 0.939859 | 0.939078 | 0.939695 | 0.944875 |
| Testing | 0.646617 | 0.781015 | 0.773026 | 0.786654 | 0.790414 |

### 4.2 Recall (以新模型結果爲表格呈現)

| 模型比較 | LightGBM | Random Forest | CatBoost | Ensemble |
|---|---|---|---|---|
| Training | 0.931937 | 0.973822 | 0.952880 | 0.952880 |
| Testing | 0.736842 | 0.842105 | 0.736842 | 0.736842 |

### 4.3 AUC ROC curve plot

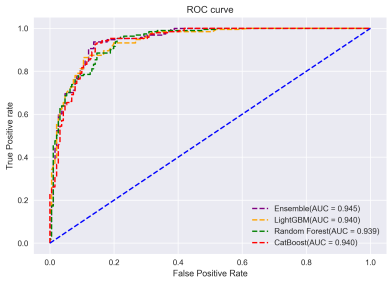


Figure 23: Training set ROC plot

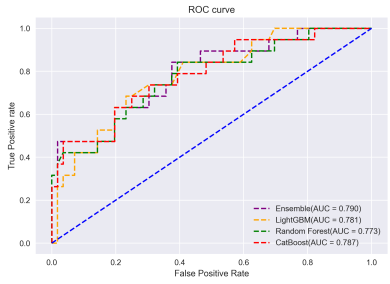


Figure 24: Testing set ROC plot

## 5 結論和未來工作

從 AUC score 的表格中，我們知道比較在純粹 Logistic regression比起 Tree Based (LightGBM, Random Forest, CatBoost) 的方法在訓練與測試集來的差一些，也利用了 10-Fold CV 去挑選有關樹的深度與葉片個數避免過擬合的問題，並且也調整在不同Tree Based的模型超參數，最終畫出 AUC ROC curve 找到最適擬合模型爲 Ensemble在訓練與測試集最好。
然而，若考慮 Recall (即爲眞實 label爲死亡，確實預測爲死亡 label的比率) 藉由我們調降 Threshold，得出在 Random Forest 值最高。因此，針對這份肝細胞癌的資料我們應當著重在 Recall Score 。

### 5.1 問題與討論

(1) 在Tree Based中，訓練集相對於測試集的 AUC Score 高很多，主要因素可能爲在測試集中，生存個數爲 57，死亡個數爲 18，測試集上資料有不平衡的問題，以至於 AUC Score 在測試集上低於訓練集許多的主因。

(2) 猜測過擬合的原因爲，樣本不多但樹的棵樹太多，於是在我們調整樹的棵樹爲10棵後得到上面的新的結果，但其實改善不大，train 大概降 0.05 左右，test 則是改變不多；可能還有我們未考慮到的地方，也許是有些我們判斷沒有 leaking 的變數，但實際上是有，而我們卻把他放入模型，這可能是未來模型改進的方向。

(3) 值得思考的是 Threshold=0.4 再調低的話，是不是合理的? 理論上對於死亡個案的眞實預測比例會變得更高，但同時也可能會增加對生存個案的預測錯誤發生率提高，必須針對研究者的思考立場做衡量。

(4) 希望有可能可以增大樣本數，能夠避免過擬合的問題

## 6 Contributions

- 資料介紹: 高念慈、林良朋
- 相關工作與變數處理 code : 高念慈
- 模型與 code : 林良朋
- 問題與討論: 高念慈、林良朋
- 書面報告彙整: 高念慈、林良朋

## References


[1] Backward feature. https://www.analyticsvidhya.com/blog/2021/04/backward-feature-elimination-and-its-implementation/.

[2] Catboost. https://catboost.ai/en/docs/references/training-parameters/.

[3] imbalanced-learn. https://imbalanced-learn.org/stable/over_sampling.html.

[4] Interpolate. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html.

[5] Labelencoder. https://pyecontech.com/2020/07/22/python_label_encoder/.

[6] Lightgbm2. https://lightgbm.readthedocs.io/en/latest/Features.html.

[7] Navaluemethod. https://ithelp.ithome.com.tw/articles/10201106.

[8] T-sne. https://mortis.tech/2019/11/program_note/664/.