111學年度第二學期
數據科學實務與創新
期中報告
111-2
PRACTICAL AND INNOVATIVE ANALYTICS IN DATA SCIENCE
Midterm Report

主題：探索學生背景因素，對於學生們是否會完成考前課程的影響

Topic：Exploring the Predictive Model of Student Background Factors on Course Completion

Abstract

透過探討學生背景因素對於是否完成考前準備，先做EDA看各個變數對於目標變數的關聯性，在經由資料預處理，得到的訓練集與測試集後，放入PyCaret做建模

組員：

M102040035 林良垵

M112040034 李祐瑄

M112040036 孫瑞鴻

目錄

1. Introduction

這份生成資料集包含了來自一所中學的學生的相關數據。該數據集共包含8個變量，包括學生的性別及種族、父母的教育程度、是否享有午餐補助、是否完成考前課程以及三個考試科目的分數，分別為數學、閱讀和寫作。

研究目的是希望透過分析學生的背景因素，例如性別、父母的教育程度和是否享有午餐補助等，以瞭解這些因素是否與學生完成考前課程之間存在相關性。

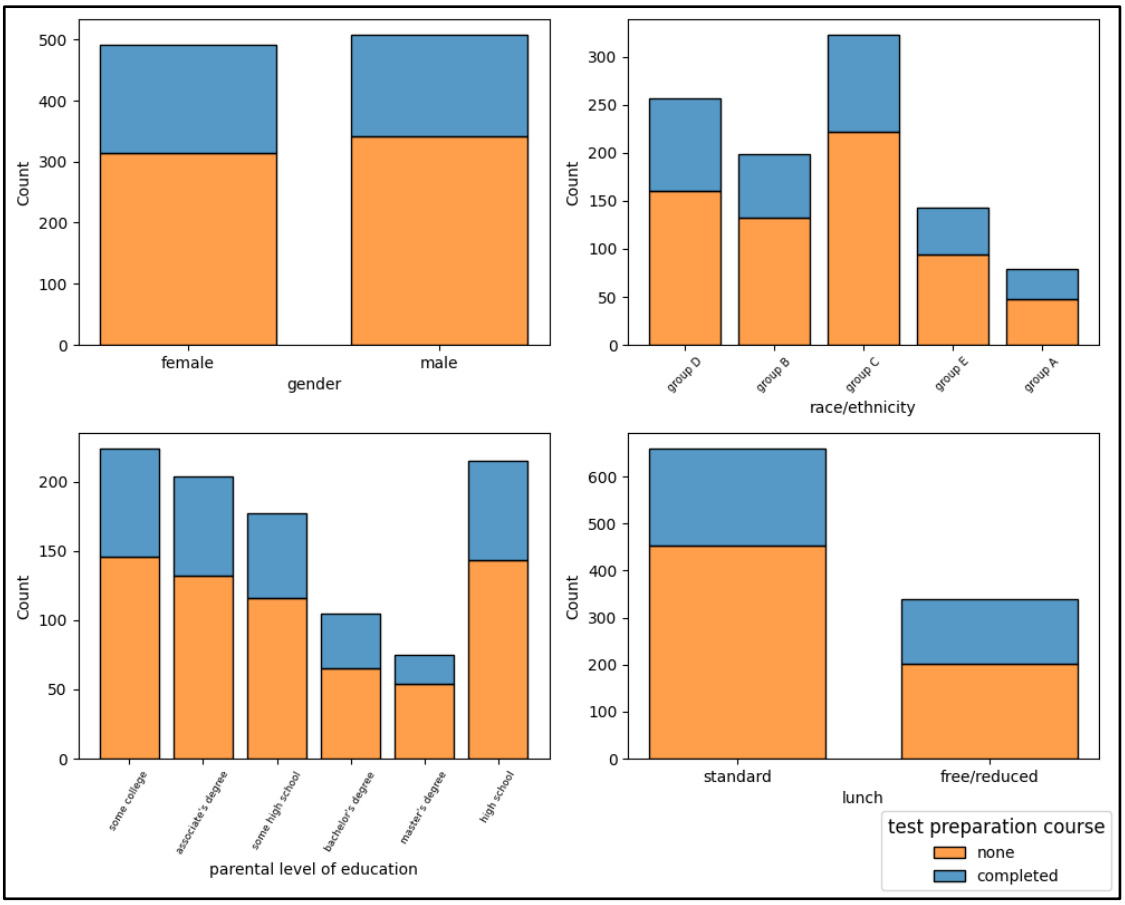
A. 資料集欄位說明

欄位名稱	欄位內容	型態
gender	male, female	類別
race/ethnicity	Group A ~ E	類別
Parental level of education	some college, high school, associate's degree, some high school, bachelor's degree	類別
lunch	standard, free/reduce	類別
Test preparation course	none, completed	類別
Math score	0~100	連續
Read score	0~100	連續
Write score	0~100	連續

▲表一

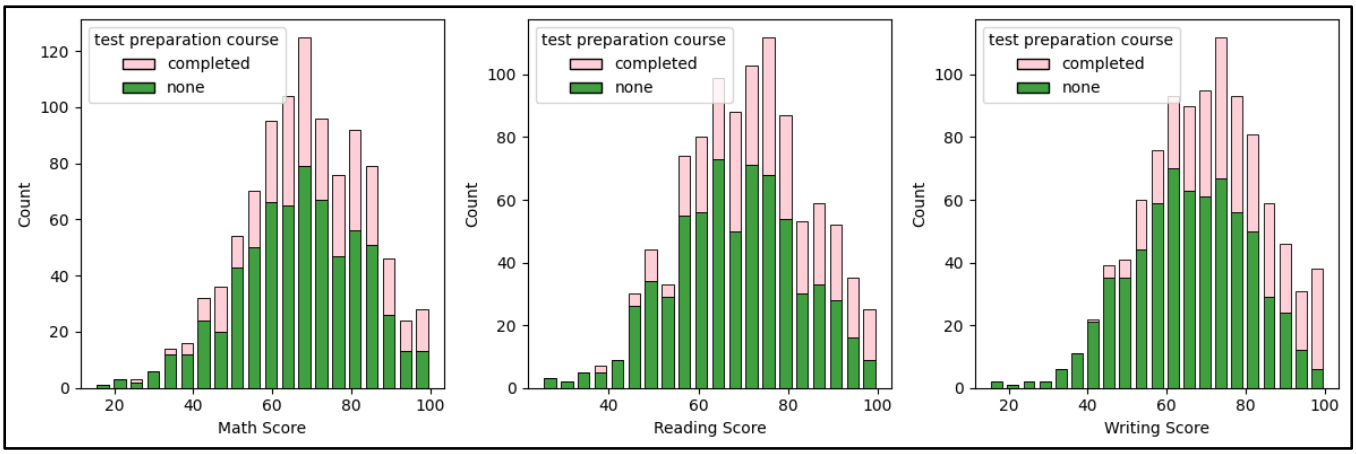
表一的「Test preparation course」，正是我們想要去預測的欄位，「none」表示沒有完成；「Completed」表示有完成考前課程，而「race/ethnicity」的部分Kaggle上沒有明確提及代號A~E分別是哪個種族，「lunch」上，standard表示沒有拿到午餐補助;free/reduce則是將全免或減免歸在一類。

B. Exploratory Data Analysis，EDA



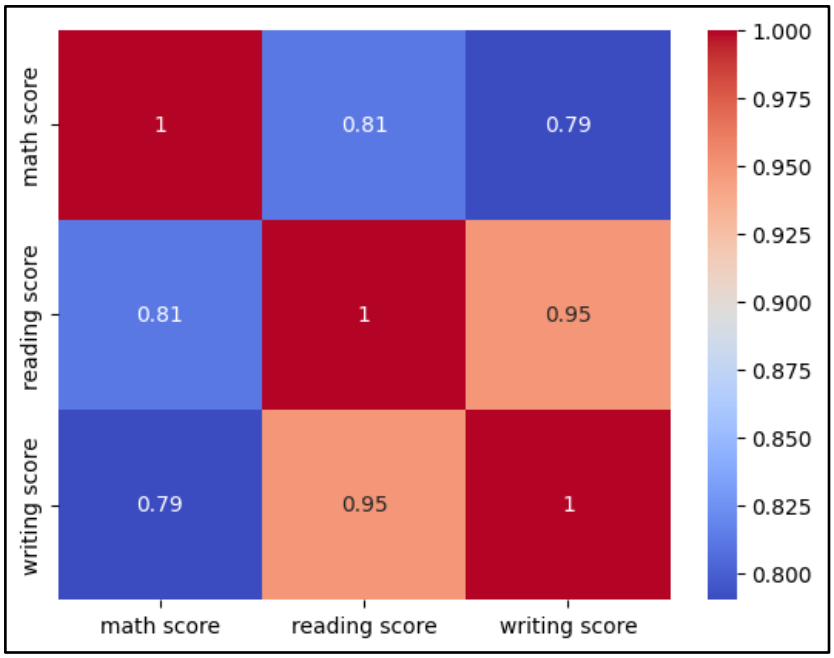
▲圖一

圖一可以看到，性別、種族、父母教育程度、午餐補助等因素，對於是否有完成考前課程的堆疊圖，這張圖橘色的部分是沒有完成考前課程，藍色的是有完成，但無論在何種因素上都沒有看出特別明顯的差異存在，意即並無因為是男生或是女生，完成課程的比例較高或是哪個種族完成課程的比例較高等。

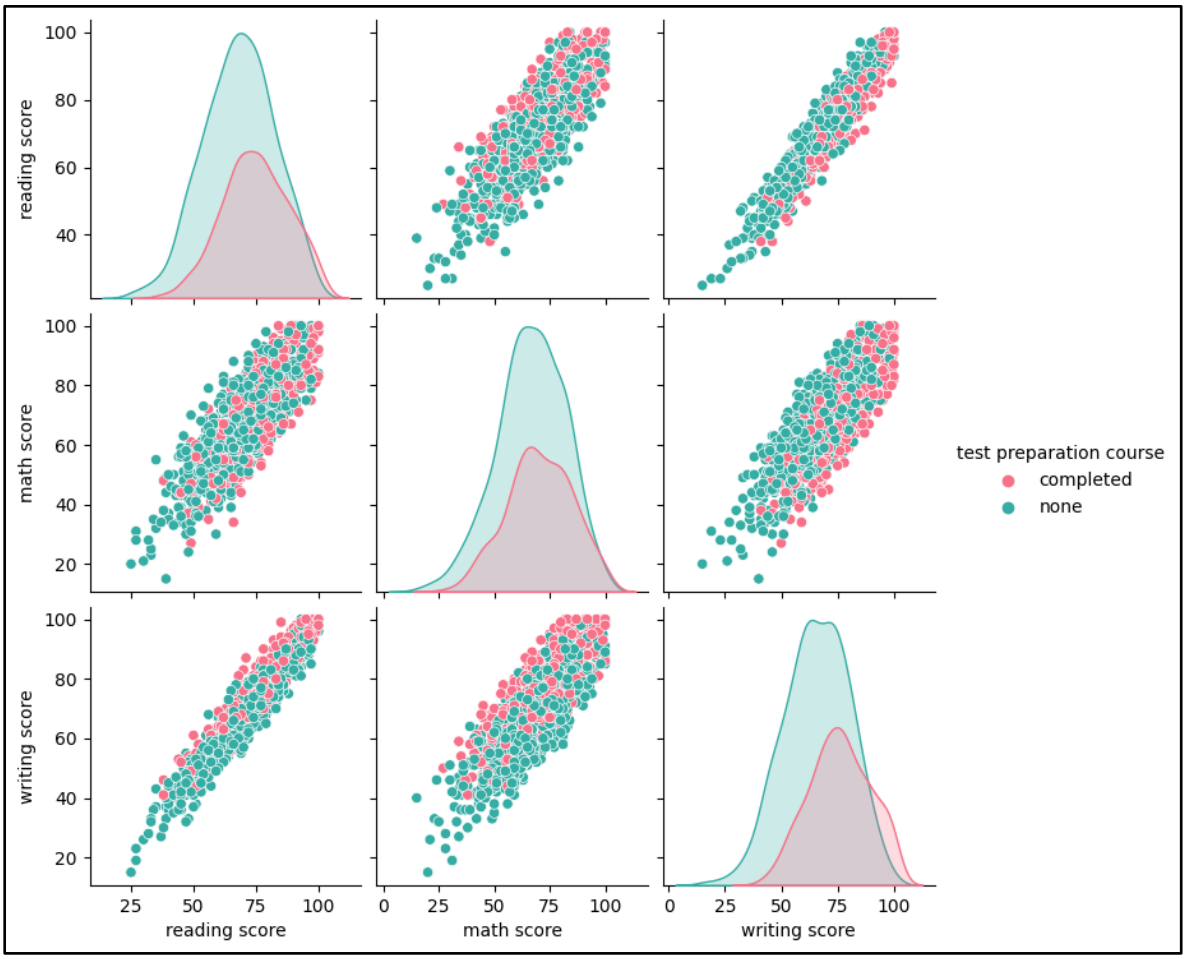


▲圖二

圖二中可以觀察三種分數的分布，大致上符合常態分佈，包含有無完成考前課程對應分數的人數，就觀察而言，可以發現無論在哪種考科上的情況，50分以下有完成考前課程的人比50分以上的人看起來少很多。

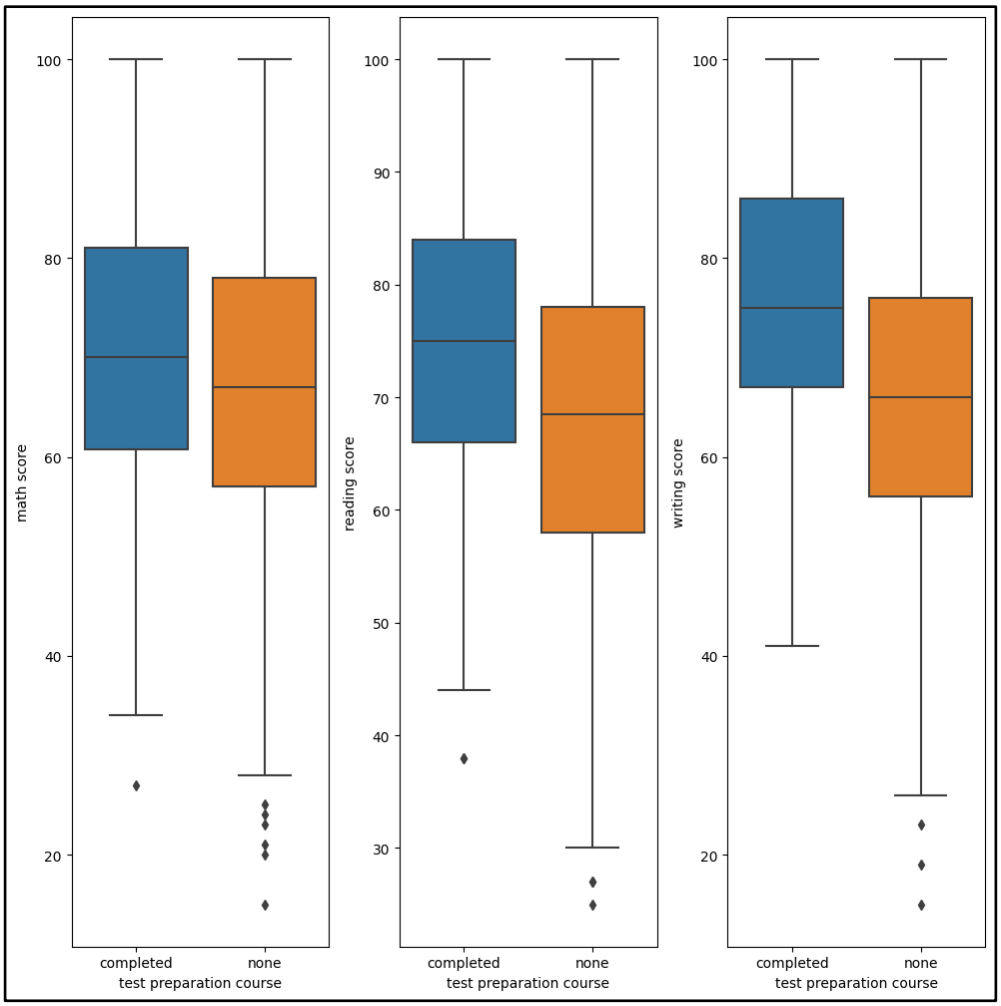


▲圖三



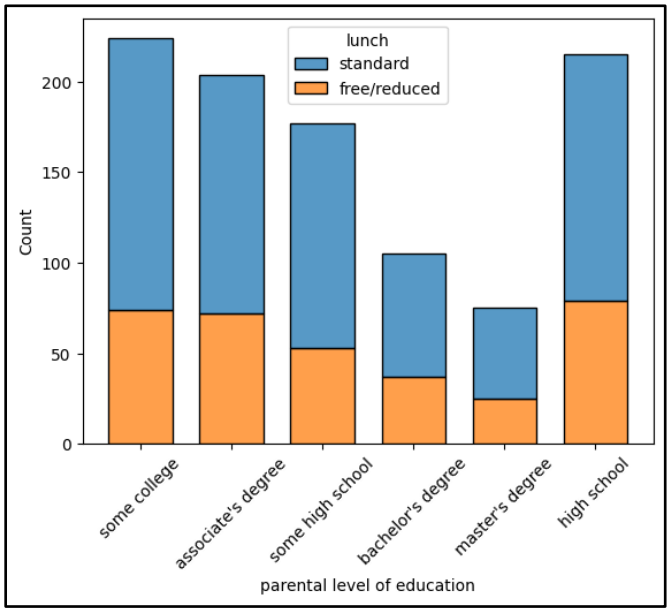
▲圖四

圖三是三種分數的相關係數圖，可以看到讀寫的相關性蠻高的，其次是閱讀與數學的相關性，最後是數學與寫作的相關性較低，但其實三個彼此都蠻高的，這可能表示說通常會三種分數一起表現得不錯，圖四則是對應的散點圖。

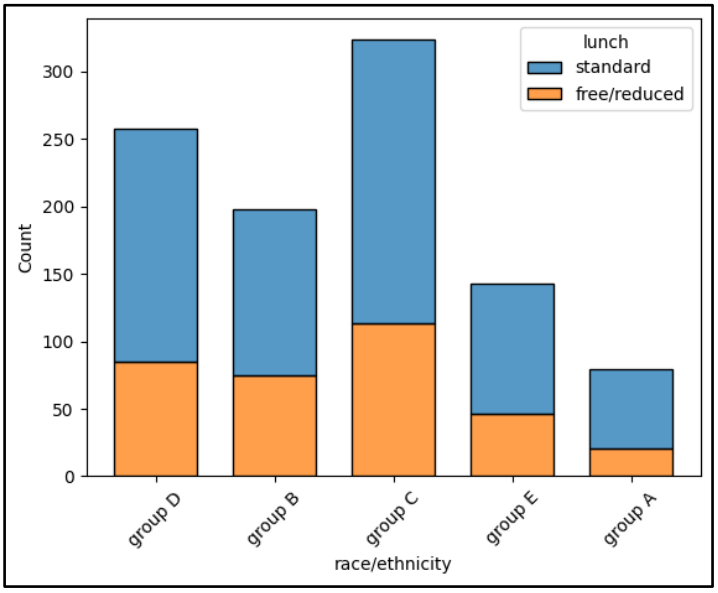


▲圖五

從圖五的box-plot，可以觀察到無論哪種分數的中位數，都是有完成考前課程的較高。



▲圖六



▲圖七

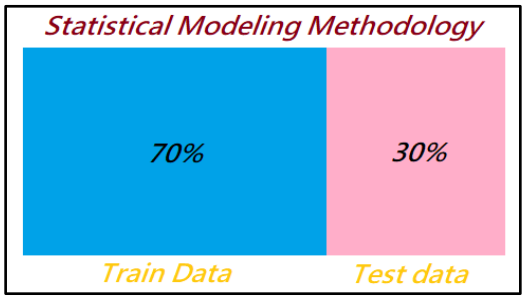
圖六、圖七是希望從父母教育程度或是種族，來去看跟拿午餐補助有沒有什麼關係，可不可以建立一些交互項，但結論是單從敘述統計觀察，無法看出哪些族群有較大的領取補助占比。

2. Preprocessing

A.確認沒有NA值

#	Column	Non-Null Count
0	gender	1000 non-null
1	race/ethnicity	1000 non-null
2	parental level of education	1000 non-null
3	lunch	1000 non-null
4	test preparation course	1000 non-null
5	math score	1000 non-null
6	reading score	1000 non-null
7	writing score	1000 non-null

B. Train Test Split



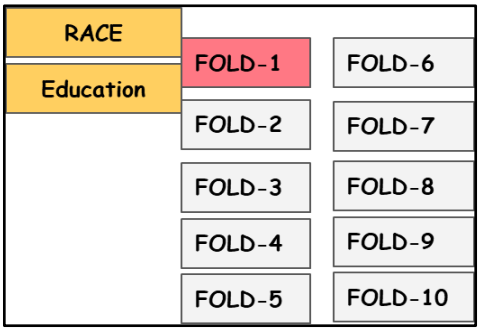
▲圖九

先做切分訓練與測試集(圖九)，以Training:Test為7:3的方式切分，random_state設定2023，以確保每次分割結果一致。

C. One-hot Encoding

原始資料集中，因性別與午餐只有兩種類別，所以採用One-hot Encoding

D. Target Encoding



▲圖十

以Race為例，因應訓練集中(圖十為例)分成10個FOLD，而FOLD-1中，以其他FOLD去估計GroupA到E中Target的平均作為編碼，在將所有FOLD得到的各個Group的估計值做平均，作為測試集的編碼值；同理Education也以此類推。

然而，在Race與Education，使用10-FOLD target encoding，使用它的優點是可以減少Overfitting的問題。另一方面，沒有做One-hot encoding是因為他們的類別數太多，若使用One-hot Encoding可能會使得維數提高，而有維數災難的問題。再者，若使用label Encoding，會有Order的問題，因為我們無從得知是否A,B以及B,C的種族差異距離是一致的，因此採用Target Encoding。

以下圖十一為經過One-hot Encoding與10-FOLD Target Encoding訓練集資料圖

	gender_female	gender_male	lunch_free/reduced	lunch_standard	math score	reading score	writing score	test preparation course	race	education
300	1	0	0	1	72	77	75	0	0.275000	0.330709
391	0	1	0	1	75	66	59	0	0.344086	0.362963
584	1	0	0	1	61	69	69	0	0.418605	0.311594
161	0	1	0	1	49	53	47	0	0.279793	0.330827
570	0	1	0	1	75	71	68	0	0.329787	0.333333
...	...	...	...	...	...	...	...	...	...	...
929	0	1	0	1	68	53	53	0	0.385965	0.321101
695	1	0	0	1	69	86	87	1	0.377246	0.333333
454	0	1	1	0	87	79	72	0	0.344086	0.431034
537	1	0	0	1	84	100	96	0	0.328125	0.349650

▲圖十一

3. Analysis

▼圖十二

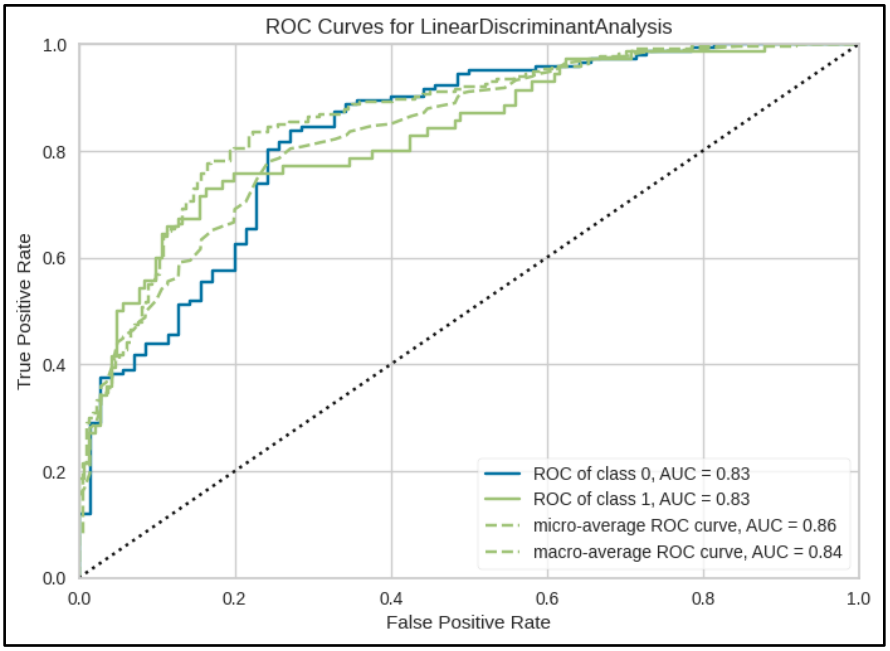
```
import numpy as np
import pandas as pd
import pycaret

from pycaret.classification import *
s = setup(data = train, target = 'test preparation course', session_id=2023)
```

利用pycaret套件，將train資料導入，預測目標為test preparation course，自動建立多個模型去做分析比較。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
lda	Linear Discriminant Analysis	0.7607	0.8064	0.5026	0.6832	0.5743	0.4164	0.4278
lr	Logistic Regression	0.7526	0.8049	0.4901	0.6706	0.5614	0.3982	0.4094
ada	Ada Boost Classifier	0.7177	0.7167	0.4254	0.6120	0.4994	0.3114	0.3229

▲圖十三

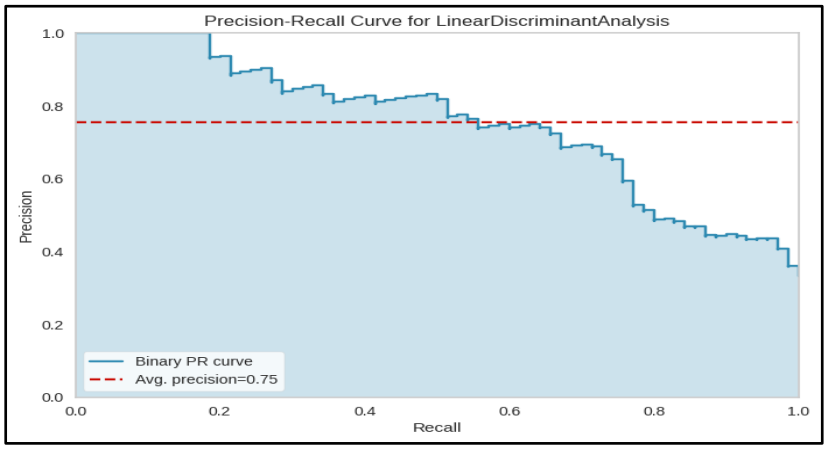
圖十三我們透過auc來排序模型，可以看出lda與lr模型看起來整體上表現差異不大，但lda略勝，因此我們選擇使用lda來當作模型。



◀圖十四

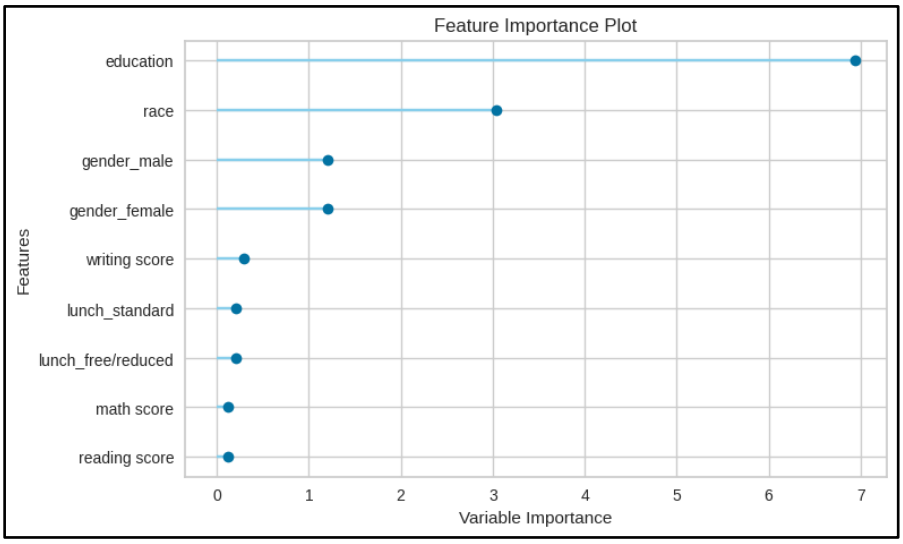
圖十四表示越往左上就代表模型性能越好，我們可以看到不管是micro-average ROC curve或是macro-average ROC curve其實AUC都有0.8左右，代表模型整體性能不錯。

3



▲圖十五

圖十五為PR曲線圖，圖形越往右上角代表模型越好，可以看到平均的precision有0.75，代表模型尚可。



▲圖十六

圖十六為特徵重要性圖，將顯示出lda模型使用的特徵及其對分類模型的貢獻。當中對模型貢獻最多的變數為education，其中會發現math score 與reading score對模型來說較不重要。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Linear Discriminant Analysis	0.7600	0.8427	0.5446	0.7439	0.6289	0.4577	0.4699

最後將測試集放入lda模型中，可以看到accuracy有達到0.76，表現尚可。

4. Conclusion

A. 比較不同特徵數的模型

若單純使用 Gender、Education及 Race的LDA，Accuracy(66.8%)且AUC(55.2%)，相較於LDA(所有特徵)，Accuracy(76.02%)，也是有不錯的解釋能力。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.6687	0.5085	0.0000	0.0000	0.0000	0.0000	0.0000	0.0520
nb	Naive Bayes	0.6687	0.5297	0.0000	0.0000	0.0000	0.0000	0.0000	0.0500
ridge	Ridge Classifier	0.6687	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0730
lda	Linear Discriminant Analysis	0.6687	0.5520	0.0000	0.0000	0.0000	0.0000	0.0000	0.0490
dummy	Dummy Classifier	0.6687	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0480

B. 未來工作

之後可能可以嘗試交互項，看看是否會改善模型的表現，但由於無法確定特徵間的相關性，目前尚未加入。

5. Contributions

Introduction	李祐瑄
EDA	李祐瑄、孫瑞鴻
Preprocessing	林良埰
Analysis	孫瑞鴻
Conclusion	林良埰
投影片&Code彙整	李祐瑄、孫瑞鴻、林良埰

6. Reference

數據集來源：https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics?resource=download&select=exams.csv

https://pycaret.gitbook.io/docs/get-started/functions/analyze

https://medium.com/@pouryaayria/k-fold-target-encoding-dfe9a594874b

https://towardsdatascience.com/introduction-to-binary-classification-with-pycaret-a37b3e89ad8d