國立臺北大學經濟學系

碩士論文

指導教授：林茂廷 博士

二一預測模型建構-以國立臺北大學日間部學士班為例

Construct A Prediction Model Of Excessive Course Failing Hazard Of the Undergraduates in National Taipei University

研究生：李冠緻

中華民國 108 年 8 月

國立臺北大學 107 學年度第二學期 碩士學位論文摘要

論文題目： 二一預測模型建構-以國立臺北大學日間部學士班為例

論文頁數：38

所組別： 經濟學 系(所) 學號： 710661105

研究生： 李冠緻 指導教授： 林茂廷 博士

論文題要內容：

現今國內將機器學習應用於學生未來是否會被二一的研究較少，若是將資料探勘應用於與教育之中，不僅可以儘早發現學生的問題給予協助，也可以知道是何種因素影響著學生被二一。本文分別利用了羅吉斯迴歸、隨機森林、支持向量機以及類神經網路模型，針對國立臺北大學日間部學士班學生未來是否會被二一進行預測。結果顯示隨機森林表現最為良好，其二一學生捕捉率（召回率）高達 81%。其中預測力的主要關鍵變數為累積專業必修排名，代表學生專業科目上的學習狀況對於學生日後是否會產生二一影響最為嚴重。本文主要貢獻為即早預測出被二一學生，並且分析出影響二一最為重要的科目為專業科目，故校方可以針對本文中未來預測為被二一的學生，加強學生於專業必修科目上的學業表現，以達到退學率降低及學習品質提升之效果。

ABSTRACT

Constructing An Prediction Model Of FMTH - A Case Study of National Taipei University Of Excessive Course Failing Hazard Of the Undergraduates in National Taipei University

by

LI, GUAN-ZHI

August 2019

ADVISOR: Dr. LIN, MAU-TING

DEPARTMENT: ECONOMICS

MAJOR: ECONOMICS

DEGREE: MASTER OF SOCIAL SCIENCE IN ECONOMICS

In Taiwan many universities expel students who fail more than half (FMTH) of their enrolled classes for two semesters. Given the populous applications of machine learning, very few are used on predicting this excessive course failing hazard. In this research, we used the transcript data from the National Taipei University undergraduate students from 2011 to 2017 to train machine learning models to predict student's FMTH.

Different classifiers have been trained on our data sets including Logistic Regression, Random Forest, Support Vector Machine(SVM) and Artificial Neural Network (ANN). The results show that Random Forest has the best recall rate which is of 81%. Further scrutiny on The study also includes variable importance measure shows that. We find the ranking of a student's cumulative GPA on major-required professional subjects ranking is the most important variable for FMTH prediction. which means the learning situation in the professional subjects of students will have the most serious impact on whether students will flunk out in the future.

Abstract contribution of this article, Our work produces the prognosis of students' FMTH possibility, which can be a valuable information for both the school and the students to strengthen learning efficacy.

目錄

圖目錄

表目錄

第一章 緒論

第一節 研究背景

現今國內外大學普遍存有督促學生學習狀況的機制，主要是以學生的在校成績作為其衡量標準，校方對於未達到標準的學生會進行懲罰，旨在希望學生不要荒廢課業。 大學退學可分為「自退」與「勒令退學」，自退為學生主動於校方申請退學，而勒令退學則屬校方強迫退學，按學校規定勒令退學將分為「非因成績退學」，以及「因成績退學」，早期之因成績退學制度為「二一退學」，即實拿學分不及學期總學分二分之一便退學，然而自教育部開放大學自主之後，陸續有許多學校對於退學標準做出調整。退學制度較為主流的為下列三者；較為嚴格的「三二退學」即實拿學分不及學期總學分三分之二便退學，較為寬鬆的「連續雙二一退學」即在校期間連續被二一兩次才會被退學，最後一項為「雙二一退學」，亦即在校期間累積被二一兩次便退學。

交通大學註冊組組長彭淑嬌指出，幾年前教育部開放給大學自主時，曾有一波廢除二一制度的聲浪，當時交大也因此廢除「單二一退學制度」，但後遺症很明顯，學生學習動力大幅減弱，校方最後又改採雙二一制度，雖陸續有學校開始廢除因成績退學制度，但目前大多數之大學仍保有著因成績退學機制；吳東陽(2018) 研究指出二一制度對於學生可以有效達到嚇阻的功用，學生在被二一後的不及格學分比例在往後的學期會有所改善，故學生是否被二一仍是一個很好的學習成效指標。

第二節 研究動機與文獻回顧

第一項 研究動機

台灣社會面臨少子化的趨勢，造成各級教育機構入學人數普遍下滑，教育部因應此趨勢於民國 102 年推動大學整併計畫，針對一縣市有兩所以上的公立大學且單

一學校學生人數在一萬人以下的公立大學推動整併；私立大學學生人數於兩千人以內，則推動退場機制。為維持學校的規模，各大院校愈來愈重視學生的退學問題；若能夠於早期預測出學生退學，以減少退學的比例，對於學生與學校將會是雙贏。

鑑於校方於學期中才能發送期中預警，本文希望能夠透過模型在學期開始前便有效的預測該學期是否會被二一，於學期初期便能給予老師或是周遭同學訊號，發揮同儕之間的影響力共同關懷學習上遭遇困難的學生；鄭媛文（2013） 研究指出教師對於學生學習成效之認知、情意及技能部分有顯著的影響，同儕教導學習策略對學生的「學習成就」、「情意態度」均有正向的影響，可見教師與同儕的影響在學習過程中扮演了息息相關的角色，透過儘早預測出需要幫助的學生，校方可以從教師與同儕方面擬出一些補救政策幫助學生。

對於學生被發送期中預警，陳家琪（2017）針對台北市立大學 103 和 104 學年度大學部的學生運用過往修課狀況進行了研究，研究顯示大一學生以及大二學生被期中預警的比例是最高的，反應了學生對於大學的生活可能存有一定程度上的不適應，且各學學生收到期中預警的原因也不盡相同，如體育學院的學生主要被預警的原因為出缺席率而理學院的學生多為成績表現上的問題。

李勁昇（2017） 也針對同間學校進行了修課習慣（學分數、時間、星期）進行分析，探討其學習成效，分別發現修課總學分數與學期平均成績大多呈負相關；下午修課平均成績皆高於上午；跨星期上課的科目（每週上課時間分二天以上），學生的平均成績顯著最低，學生成績表現不好不僅學校會影響學校的學生人數，對於學生心理狀況與人際狀況也會有所影響，根據李易倫（2015） 研究指出學業成績的自我覺察與人際關係具有正向的關聯性，以及大學生的學業成績與課堂焦慮具有負向關聯，若能早點給予學生成績上的幫助，除提升成績外也可以降低學生的心裡壓力。

第二項 機器學習的應用

目前對於教育的資料探勘研究比起金融、醫療領域來說相對較少，將資料探勘運用於教育領域之中稱為教育資料探勘，教育資料探勘主要可以透過萃取影響學生學習的因素，加強我們對於學習以及教育的理解。 Abu-Oda and El-Halees (2015)表示目前高等教育遭遇許多問題，導致教育機構逐步遠離了實現重視教育質量的目標，Baepler and Murdoch (2010)進一步指出大多數的原因來自於校方與學生之間的資訊落差；校方無法獲得足夠多的信息來為學生提供合適的教育，若校方能夠透過數據探勘預測學生的類型，將可以使高等教育機構以個別化的方式做出更好的決策，擬定教育時可以為學生採取較為客製化的計劃，使得教育機構能夠更有效地分配資源和人力。

Abu-Oda, El-Halees (2015)也運用了 ALAQSA 大學計算機科學系的成績單與高中成績來預測輟學的學生，希望透過預測結果針對教育有較好的理解；在研究中針對預測使用兩種演算法；分別為決策樹模型以及 Naive Bayes 模型，分別得出 98.14%與 96.86%的準確度，達到良好的預測結果，而使用較多演算法預測輟學相關的文獻有 Kotsiantis, Pierrakeas and Pintelas (2004)對於與學生的學習成效進行了相關的預測，將機器學習應用於 Hellenic Open University 遠程教育預測中；預測出哪一類型學生輟學的可能性最高；使得遠端導師可以採取預防措施減少學生的輟學率，Schaffer (1994)提到由於每個機器學習的演算法都各自擁有一些偏差值，也就代表說某個演算法在 A 領域中表現良好，很有可能在 B 領域表現是不佳的，故必須對各個不同的演算法做出比較。

在 Kotsiantis, Pierrakeas and Pintelas (2004) 研究中分別提出了六種演算法預測輟學，其中包括；決策樹、羅吉斯迴歸、Naive Bayes、 K-nearest neighbor、人工神經網路、支持向量機，其中 Naive Bayes 演算法表現最佳，其平均預測準確率為 70.51%， Cortez and Silva (2013) 也使用了隨機森林、支持向量機、人工神經網路以及決策樹對中學生的數學科目以及葡萄牙語科目進行了成績的預測，研究顯示在已

知過去成績下此預測可以達到很高的準確率，這與 Kotsiantis (2004)中的結論相互呼應：學生的現今成就表現受過去的成績表現影響很大，儘管如此預測力較高的模型通常也包含了以下的其他因素，如：學校相關(例如：缺勤人數、選擇學校的理由、額外的教育)，人口統計學(例如：學生的年齡、父母的工作和教育)和社交(例如：與朋友外出)變數仍存在很大的影響。

另外 Dekker, Pechenizkiy and Vleeshouwers (2009)於研究中提到輟學學生中有一類型特別的學生，稱之為風險類學生，此類型的學生特點在於，有高機率是不會被退學的學生，卻因種種原因被退學；也就是說校方必須提供更多的資源在他們身上，他們才能免於被退學，研究中透過高中以及大學的成績資料建構決策樹模型以提前預測出此類型的學生，準確度高達 75%至 80%之間，這項研究將有助於學生與老師共同改善學生的成績表現，使得校方可以降低學生的輟學比例，綜觀上述案例不難發現，決策樹於教育資料探勘中非常受歡迎， Baradwaj and Pal (2012)提出因為它們產生的分類規則比起其他的演算法分類更為直覺，在他所製作的文獻中也在決策樹運用於分類學生的成績表現中得到不錯的預測效果，依變數重要性提取出的幾個重要變數可以有效的預測學生在期末考的表現，有助於提早的識別需要幫助的學生，以利老師提供適當的諮詢與建議，綜合上述文獻回顧中，本文將使用上述所提及較常使用的演算法來進行二元分類的預測，包括羅吉斯迴歸、人工神經網路、支持向量機以及隨機森林。

第二章 資料說明與觀察

第一節 資料簡介

本文所使用原始資料為大學部 100 至 106 學年成績單資料，含有 1 萬 717 位學生資料，資料來源為國立臺北大學校務研究辦公室，成績單中含有少許 100 年以前入學之學生不完整資料，成績單欄位有以下幾者：系級、學號、姓名、學期成績、

科目代碼、科目名稱、學分數、開課系所、修課人數、班別、授課老師、學年、學期、必選修類別（必／選／通）、授課語言、上課時間及教室。

第二節 資料處理

由於成績單中所含有的資訊量過於龐大與雜亂，無法直接納入模型進行預測，本節我們將對資料的特徵進行改建並依序觀察與介紹。

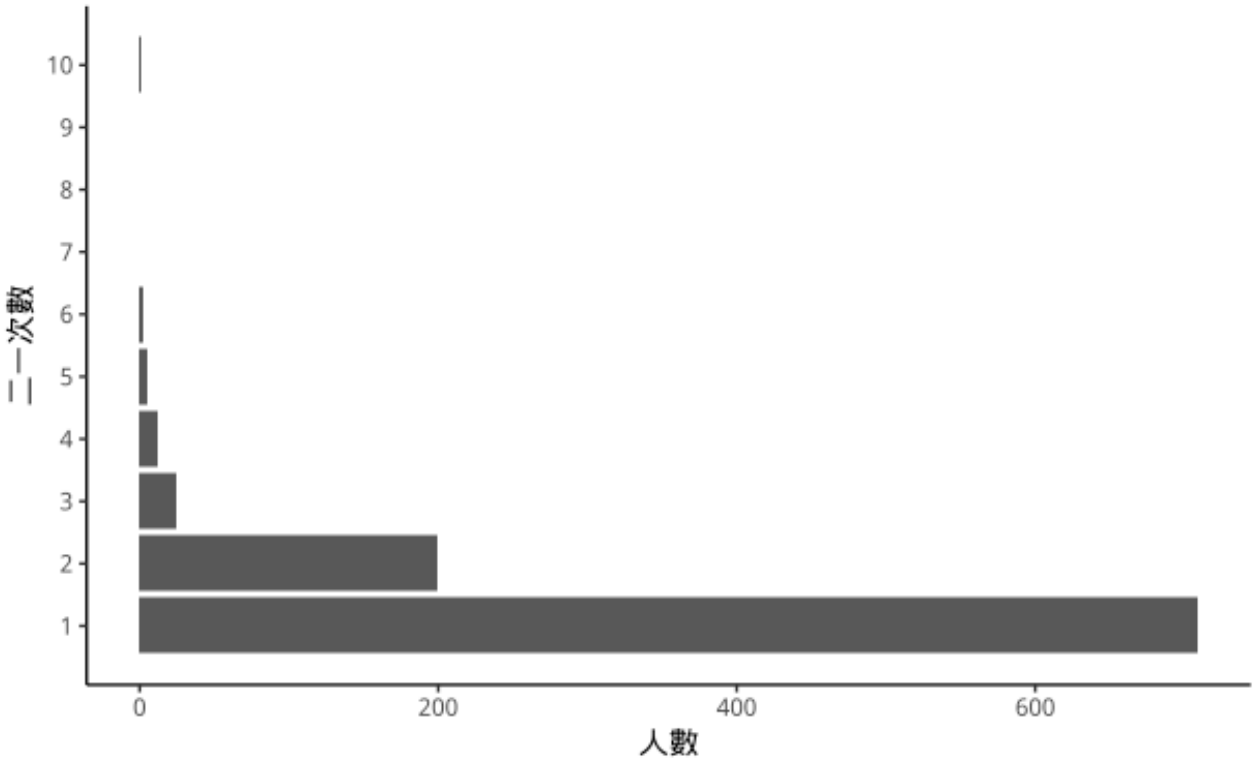
第一項 預測變數二一狀況觀察

本節討論預測目標變數，學生二一狀況，並從二一頻率觀察及不同入學年之二一狀況觀察兩個面向討論。

二一頻率觀察

圖 2-1 中顯示了成績單資料中學生的二一狀況，成績單中共有 10717 位學生，曾有二一記錄的學生人數有 952 位，佔學生人數的 8.8%，與沒有二一記錄的學生人數相比其比例相差甚大，為典型的不平衡資料。其中被二一壹次的人數最為多筆，共有 709 位，被二一兩次的人數有 199 位；三次以上的人數則有 44 位。二一壹次及兩次相差人數不小，有一部份可能是本校為雙二一制度，依據校內規定一般在校生若被二一次數兩次後便會強制退學，故有壹次記錄的學生會努力避免第二次二一。然資料中依然出現部份學生被二一次數超過兩次，這因為雙二一制度並不適用在僑生或是身心障礙學生。

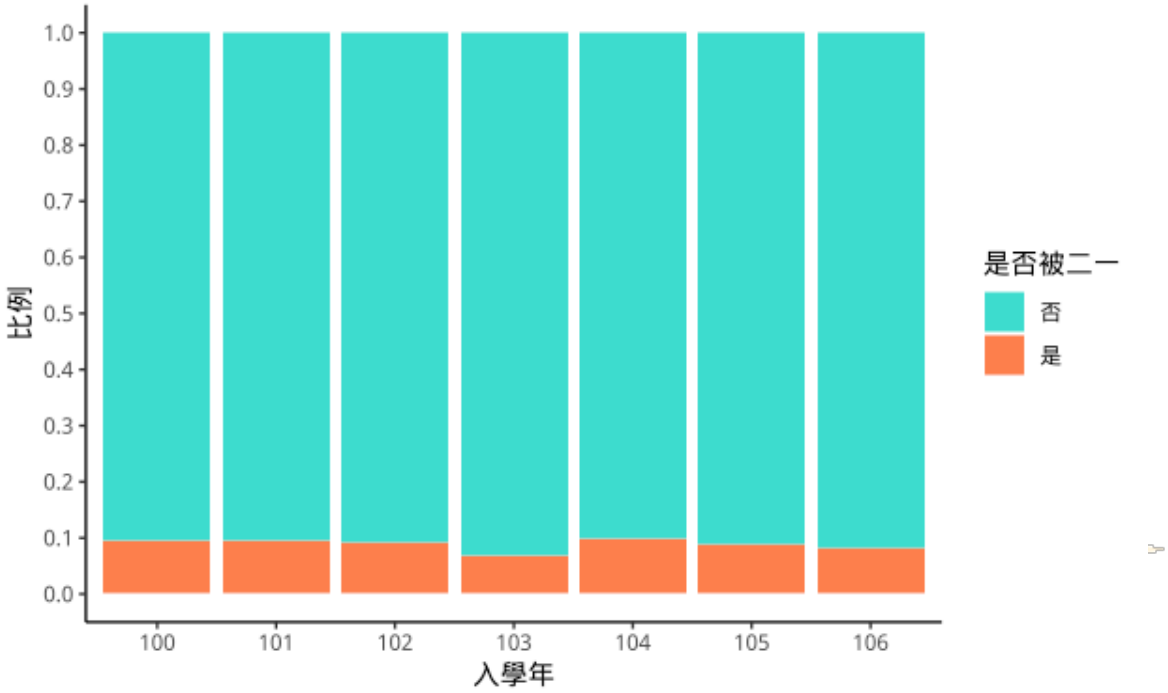
圖 2-1：學生二一頻率圖



不同入學年之二一狀況觀察

不同學年課程結構及教師評分方式可能有變，故有必要觀察不同入學年學生被二一的比例是否有所差異。圖 2-2 顯示了不同入學年學生有/無二一記錄的比例，由圖中可知各入學年的二一學生人數比例上相對隱定，皆約佔一成，代表著入學年對學生被二一 可能不具有預測力。

圖 2-2：入學年與學生二一關係圖



第二項 二一預測特徵變數

接下來我們一一探討其他可能有助於預測學生二一之特徵變數，主要變數架構依據來自於成績單資料。我們自原始成績單中各欄位改建較為有用的特徵，除了過去成績單特徵變數有助於預測未來學期同學是否會被二一，另因本研究擬於每學期初去預測學期結束後個別學生被二一的可能性，預測學期當期的選課安排也在特徵變數選擇範圍；下小節將依「預測學期以前」的訊息以及「預測學期初」的訊息分別介紹各類特徵與二一狀態的關連來找出有預測價值的特徵變數。

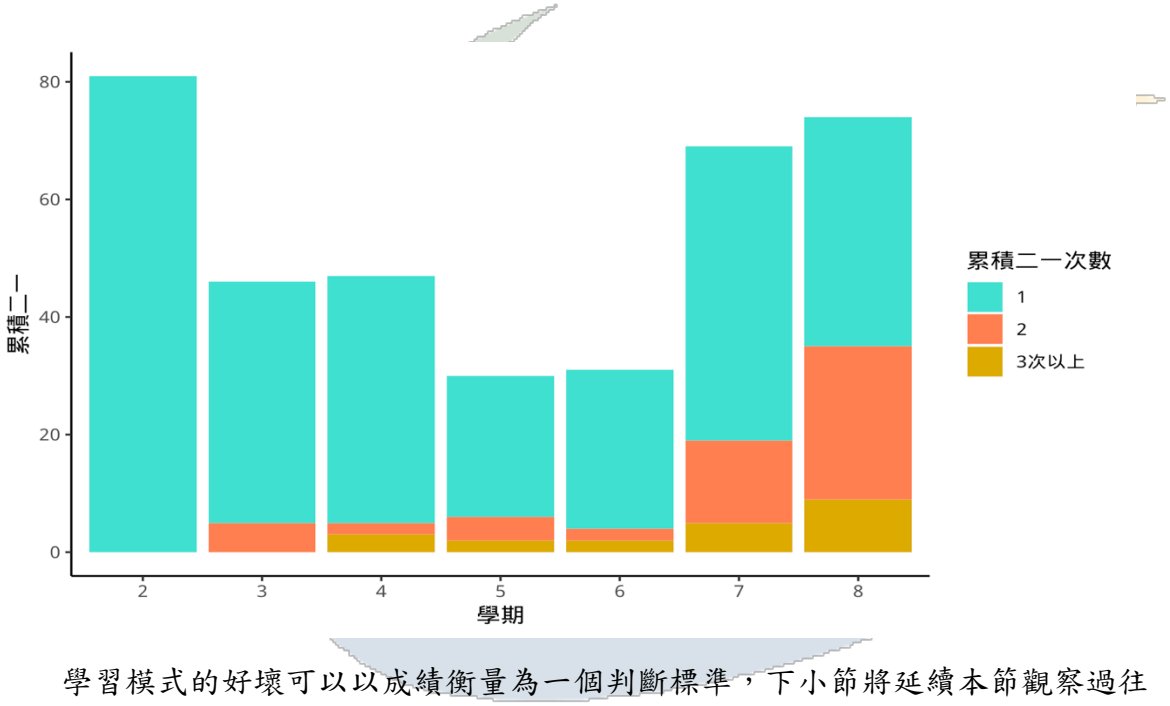
第三項 預測學期以前的訊息

由二一觀察頻率小節中，可以清楚看到大部份學生於發生一次二一後下次再度發生的頻率不高；二一次數兩次以上相對一次來得少，在此想更深入探討了被二一的學生以往的累積二一狀況為何？以及於哪些年級是學生被二一最常發生的時間點？

累積二一

圖 2-3 為被二一的學生中，累積至預測當期前的二一狀況，橫軸數字代表第幾學期，圖顯示，學生於一年級時發生被二一的狀況是最多的，故「年級」會是預測第一次二一的重要特徵。此外，要有第一次二一會有第二次二一可能，圖中顯示第 2 次多發生在大四，故「大四」與「累積二一一次」有助於預測大四第二次二一.綜合以上討論，「年級」及「累積二一次數」都是重要的預測二一特徵變數。

圖 2-3：學生累積二一與學期圖



學習模式的好壞可以以成績衡量為一個判斷標準，下小節將延續本節觀察過往成績面向的表現探討學習模式良好與不適應的學生於未來二一的狀況

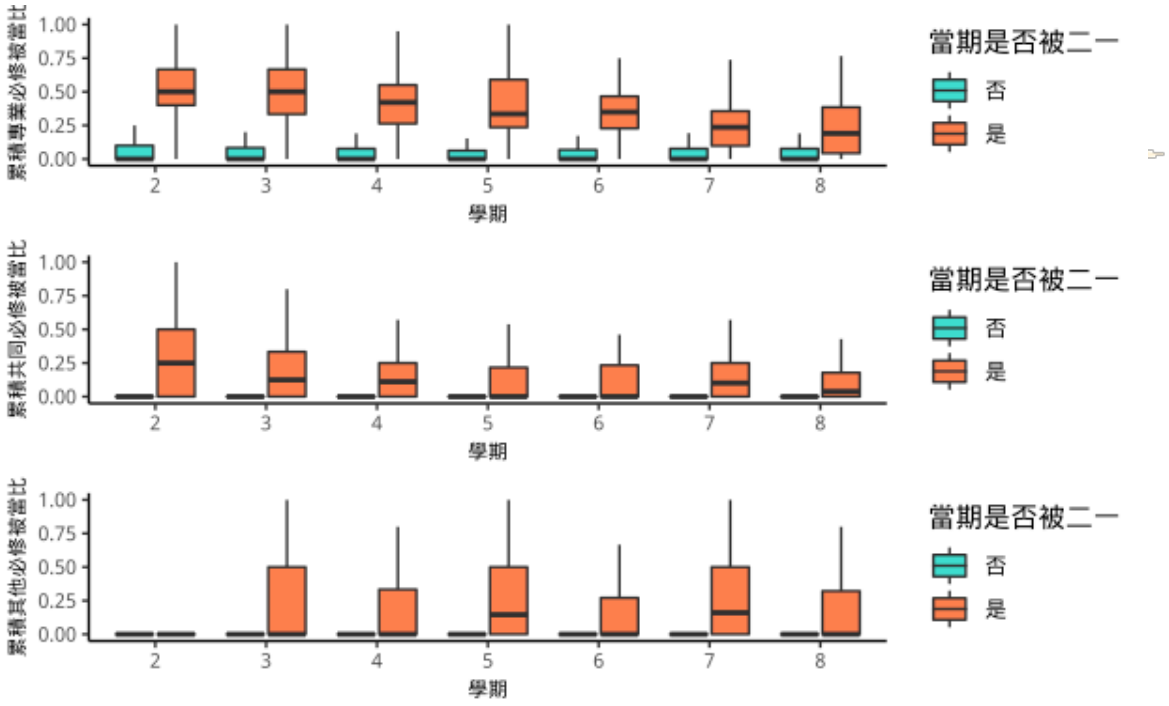
各類型課程累積被當比

大學課程類型大類可分成「必修」、「選修」及「通識」，其中必修[1]又可細分成專業必修、共同必修、其他必修三類，總共 5 類。 而在這五類課程我們去統計學生到前學期為止的在各類課程被當狀況，做為課業適應指標。這裡我們所使用

[1] 專業必修1指得是該生所屬學系的必修，其他必修為雙主修、輔修或是修習教育學程學生另外的必修課，共同必修為本校之共同必修，涵蓋體育、英文、英文聽講、歷史以及國文。

的適應狀況特徵為：學生至該學期為止在各類所修的總課程數目中被當掉的比例 — 此比例越高表示該學生在該類課程適應越不良。以此，我們建構了五個累積被當比例：累積專業必修被當比例、累積共同必修被當比例、累積其他必修被當比例、累積選修被當比例以及累積通識被當比例。採用比例定義的原因為，每人在各類的修課數受到系級、個人是否雙主修、輔修、教育學程或是不同入學年所影響，為能夠排除立足點不同的情形，故採用比例制。

圖 2-4：累積必修被當比與學生二一關係圖



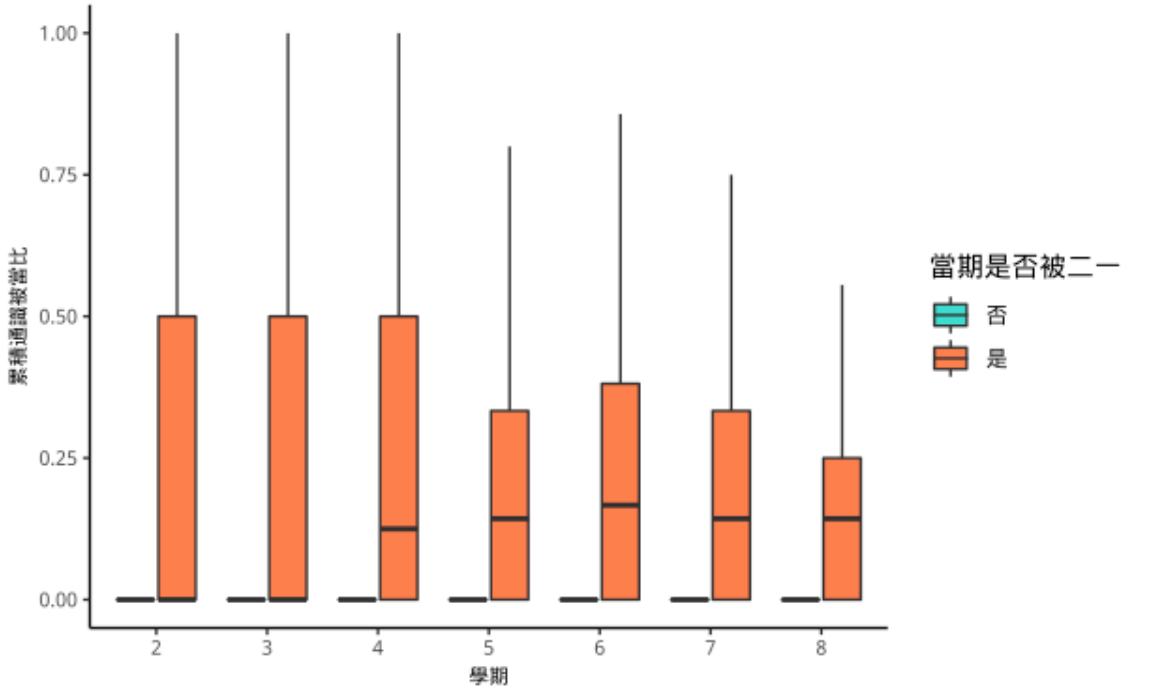
累積專業必修被當比可以看出一位學生在本系專業領域的修課狀況，由圖 2-4 可知於專業領域上，當期沒被二一的學生累積專業必修被當比明顯比當期被二一者低。由此可知一位學生若在以往的專業科目上表現良好，那麼可以預期到，他在未來被二一的可能性是較低的。

於累積其他必修類別中，大一時兩類學生在該類的累積被當比第 75 百分位數皆為 0，其主要原因為大一還不會申請如輔系、雙學位、教育學程，故無其他必修

被當可能。而在大三至大四時，此特徵在兩類學生有明顯差異，有可能此類特徵可以捕捉到，學生於外系專業上的不適應，因而導致未來被二一可能性提高。

必修類裡最後一類為累積共同必修被當比，所捕捉的是同學對於學校共同科目被當的狀況，而全校共同科目必須顧及全校各系的學生程度，所以通常會是一門相對不容易被當的科目，若此類型的科目被當，其給出的訊號很可能為學生對大學生活的不適應，包括時間控管、人際關係等等非學習上的因素。圖 2-4 中清楚呈現，當期沒被二一類型的學生此類被當比第 75 百分位數皆為 0，代表此類裡大部分的學生，對於大學生活適應程度不算太糟，而當期被二一類學生被當比的第 75 百分位數高於 0，顯示此特徵與未來被二一可能有正向的關係。

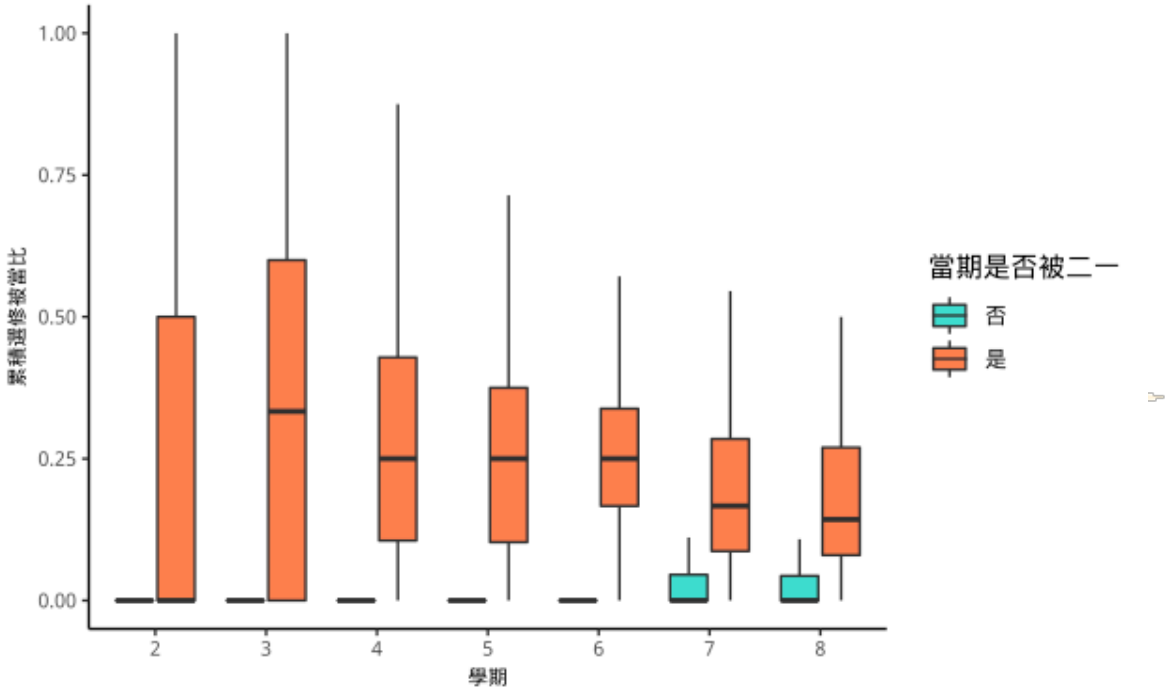
圖 2-5：累積通識被當比與學生二一關係圖



通識課程旨在培養學生五大通識素養，包括人文、科學、民主、倫理與宏觀，從累積通識被當比可以判斷一位學生對於課外的素養培養的重視程度。圖 2-5 顯示未來容易被二一的學生，對於主修專業知識以外的探索，相對於沒被二一的學生，可能較不重視而容易被當。然沒被二一類學生與被二一類學生此被當比例的中位數

差異不大，主要因為此類課程難易度較低而較不易被當。除此之外，通識科目有高度的選課彈性，通常成績較差的學生會趨吉避凶，選擇較容易的通識課程。

圖 2-6：累積選修被當比與學生二一關係圖



選修課程為一門更深入於專業領域的學科；若學生對某於專業領域中感到興趣，便會選修相關的課程，加強自己不管是外系亦或是本系的專業知識。可以預期，若學生在專業必修中或是其他必修的學習狀況不佳，那有很大的機會其累積選修被當比也會較高。圖 2-6 中顯示了兩類學生在此特徵的差異，從中可以看到兩類學生的第 75 百分位數差異很大——若學生選修被當比很高，此學生於未來被二一機會也較高。

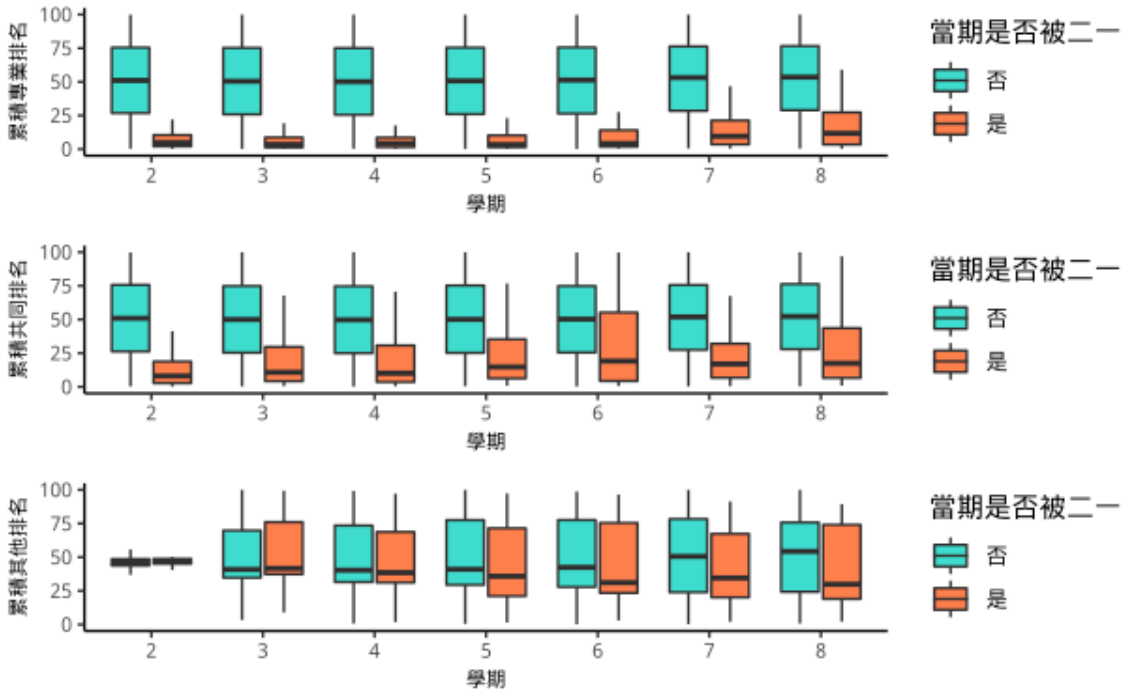
綜合上述分析可知，兩類學生在累積被當比變數群中的差異顯著，說明了它們是合理的預測二一特徵變數。

各類型課程累積成績排序

上節中討論了必修、選修課程以及通識課程的被當狀況與兩類學生之間的關係，並且知道到被二一類學生會在較有選課彈性的科目上趨吉避凶，故只看科目有沒有被當並無法完全抓到學生學習狀況，有必要進一步觀察其在課程上的排名，才能更精準的了解學習狀況。

與上小節一樣我們針對不同的課程類別進行分別討論，並計算各類的累積成績排名，其計算方式如下。先計算該類累積平均成績，方法為：計算累積至前一期某類別（如：累積專業必修）的課程總成績，除以累積至前一期某類別（如：累積專業必修）的修課數，得出累積平均總成績；再以此成績與同班的同學進行百分制排名，最高分數同學會在此排名指標中對應到 100，最低分則為 0。

圖 2-7：累積必修排名與學生二一關係圖

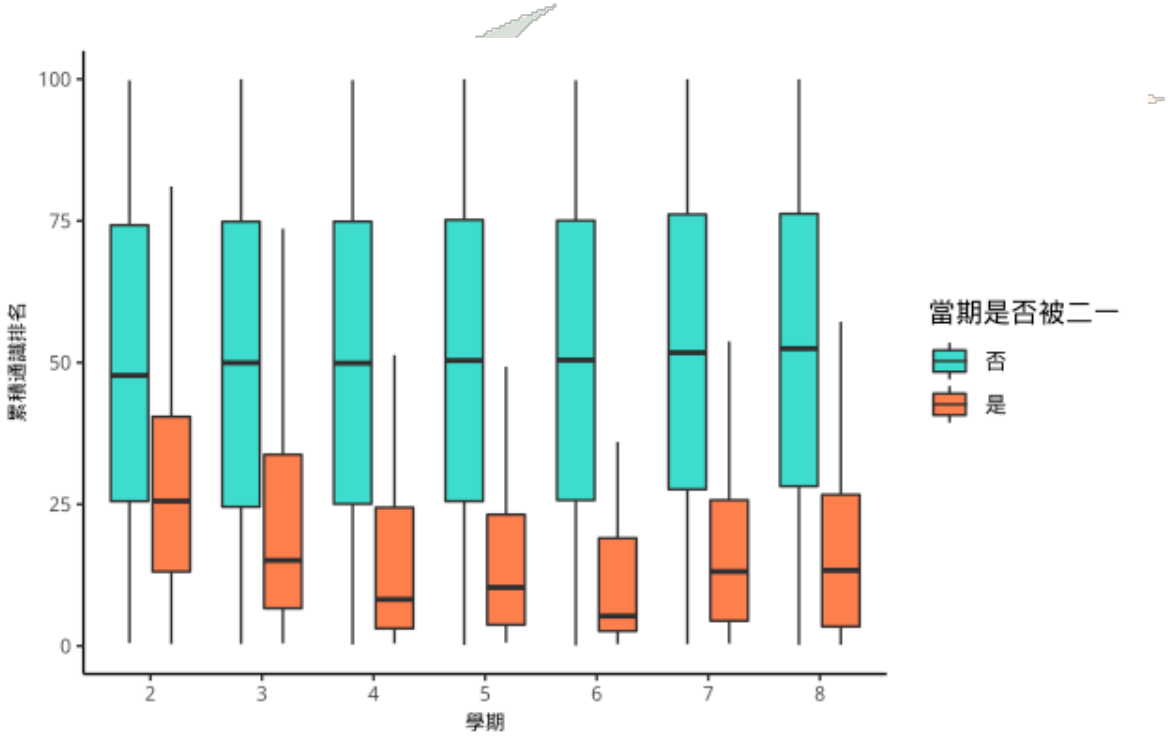


首先探討累積專業必修排名與學生二一狀況在各學期的關連。由圖 2-7 觀察，被二一的同學於以往的專業知識排名表現上低於沒被二一生許多；二一生幾乎 7 成 5 以上其過去排名皆低於 25%。

接著探討累積共同必修排名，基於共同必修負擔較輕，其排序可代表著學生在學適應狀況，圖 2-7 顯示，被二一的學生於以往的在學適應狀況（以排名判斷）會低於沒被二一生，此與前小節相互呼應。

最後為累積其他必修排名。此特徵較為特別，修此類別課程的學生人數相對其他兩類較少（因二一生不太會去輔修/雙主修），故可以發現圖 2-7 中，兩類型學生的排名中位數沒有比前幾類必修課程差異還大，不過還是可以看出二一生與非二一生之間存有差異。

圖 2-8：累積通識排名與學生二一關係圖



由圖 2-8 可知，被二一的學生類型通常為過往較不認真培養通識五大素養的學生，而發現此處中位數的差異比起累積通識被當比更為明顯，原因為此類課程屬於較不易當人的類型，且其選課彈性也比其他類型的科目都來得大，中位數的差異於被當比中較無法被觀察到的——也就是說二一生在通識課雖與非二一生一樣不容易被當，但成績上卻有明顯差別。

圖 2-9：累積選修排名與學生二一關係圖

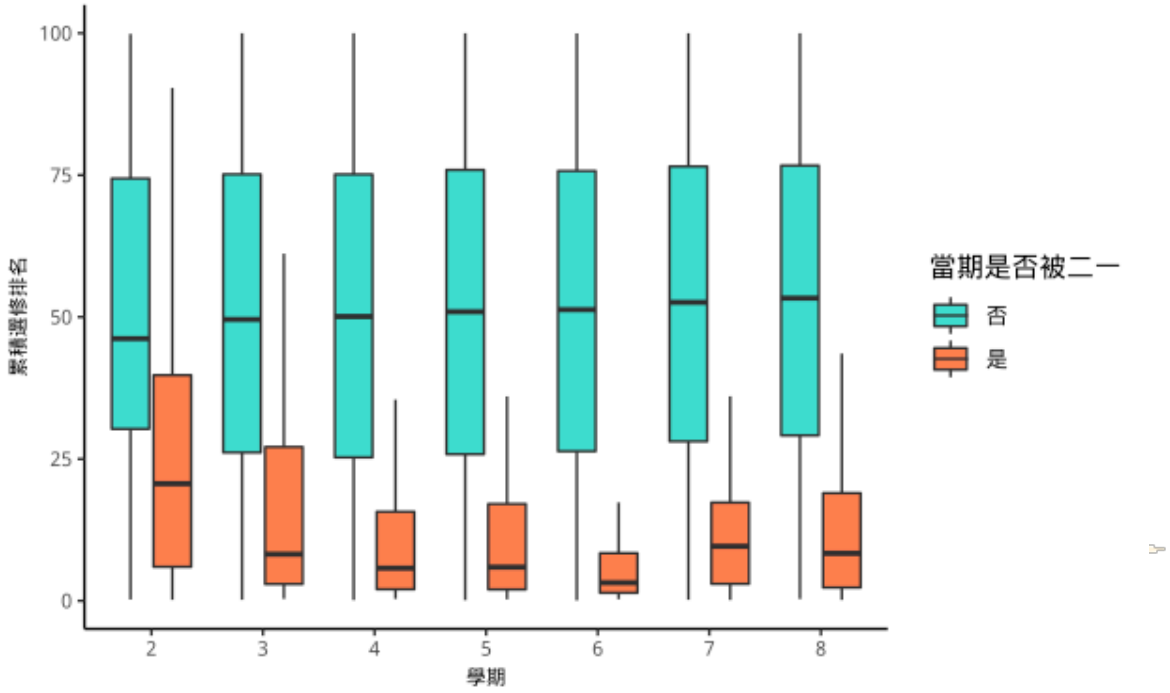


圖 2-9 為累積選修排名，兩類學生差異很大，可以作為一個很好的預測變數。

通識課程與共同必修課程這類型的課程，相對於其他課程較不會出現當人的狀況，而導致我們在累積被當比小節中，累積被當比例的中位數二一生與非二一生皆為 0。而在本節成績衡量當中，成績排序卻可以進一步補捉到兩類學生的差異性，有助二一預測是合理的預測特徵變數。

第四項 預測學期當學期訊息

由於本研究目的為預測當學期結束前該生二一可能，所以學生當學期所排定的學業課程繁重度也是重要的訊息。

在學期尚未結束前我們能夠得知關於該學期特徵的資訊僅有該學期必修數、選修數、通識數此類修課狀況特徵，為了能夠有效衡量學生在該學期修課的繁重程度，衡量方式將依據不同類型課程而有所差異。此處必修課類中僅拆成兩小類，「專業與其他必修」以及「共同必修」，主因為其他必修類別課程通常為外系的專

業必修課程，其帶來的繁重程度與專業必修是相同的，故在此節中，歸屬同類型課程。

專業與其他必修課程的衡量特徵是，以當學期修了多少專業必修課程與其他必修課程總合，除以學生於本系畢業時應修的專業必修課數，此特徵衡量的概念是，學生在當學期選的專業必修課數（包含本系與外系）佔了多少畢業時所需的必修課數，此值越高代表，當學期的專業必修課程（包含本系與外系）繁重度越高。

共同必修衡量特徵是以當學期共同必修修課數除以十一，其原因為校方規定畢業時必須修完上下學期共十一門共同必修課，分別為國文上下學期、英文上下學期、英文聽講一學期、歷史上下學期以及體育四學期。

通識類課程衡量方式是以當學期通識修課數除以十二，統一除以十二原因為，校方規定最低畢業門檻為修滿上下學期共十二門通識課。

選修類型課程的衡量較為特殊，因各系所要求的畢業學分不同、必修數也不同，所以算法為，以當學期選修修課數除以，同系所同學於當學期所選修的選課數中位數。

圖 2-10：修課比例與學生二一關係圖

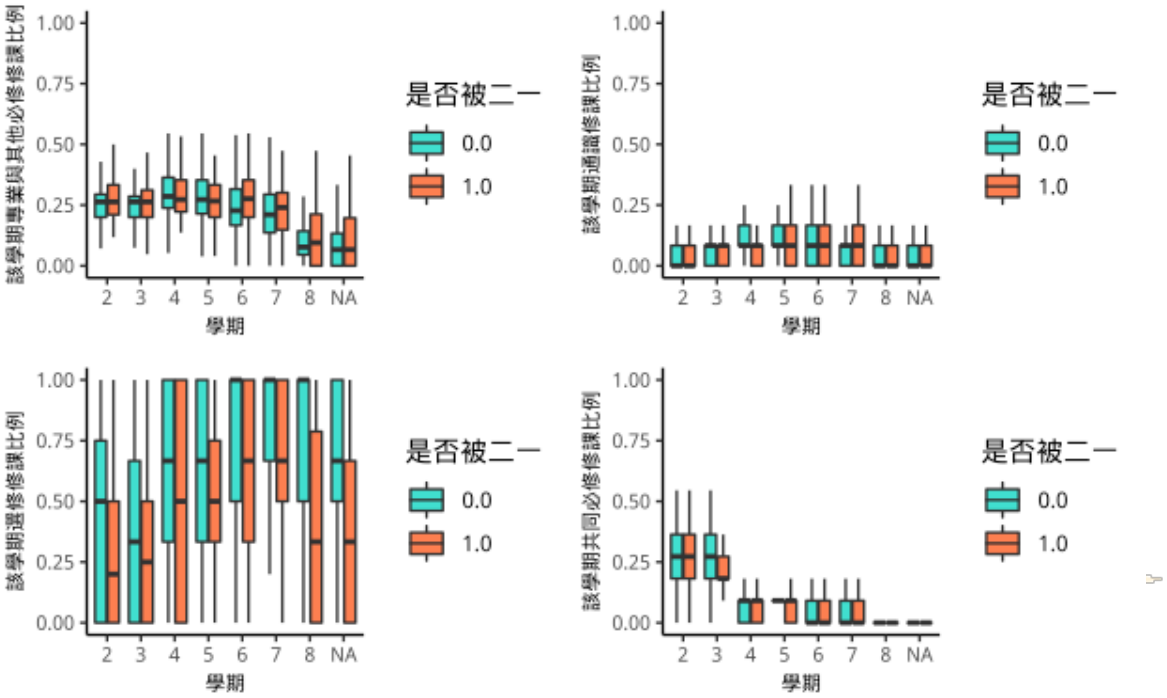


圖 2-10 中，除了該學期選修修課比例可以看出中位數擁有較明顯的差異外，其餘變數相對來說看不出趨勢為何，故推論修課的繁重程度，除選修修課比例外，並無法作為一個好的預測變數。

群聚指標解釋變數

梁玲（2016）研究顯示，同儕對於學生學習也存有很大的關係，此節將討論同儕面向所帶來的影響。首先，過去中曾有過必修科目被當的學生，他在未來選課時，會因課堂衝堂或是擋修的問題，導致這位學生與班上其他同學所安排的課表間產生重疊度降低的情況，而有前幾小項分析得知於以往必修科目表現較差的學生，在當期往往容易被二一，故若能夠加入某個特徵可以捕捉此類情形，對於預測二一相信可以達到不錯的結果。

除了上述原因外，另一個原因則顧及心理層面，部分成績表現不理想的學生，在安排課表時，會傾向選擇不與班上同學見面，藉此逃避同儕之間所帶來的壓力，導致這類型的學生課表與班上同學的課表重疊度也顯現出較低的情形。而此心理層

面的資料較為敏感，也取得不易，故若能藉由加入某個特徵，試圖捕捉學生心理層面的狀態對於預測學生二一也有良好的幫助。

綜合以上兩者原因，在此建立一個新的特徵，「群聚指標」，此指標是用來反應學生與同班同學間交集的頻率；其定義為，學期所修習的課程中，每門課裡分別有多少位同班同學一同修課，將其依依加總得到此指標。

藉由每一學期的群聚指標可以得到累積群聚指標，而站在學期初預測學生當期是否會被二一時，我們可以藉由選課系統，知道每位學生的選課狀況。故在興建累積群聚指標此特徵時，特徵期數是以累積至預測當期（包含預測當期）的群聚指標建構。又因累積群聚指標會受到不同班級課程上的安排有所不同；所以分別以各班平均以及標準差，標準化此指標。

圖 2-11：累積標準化群聚指標與學生二一關係圖

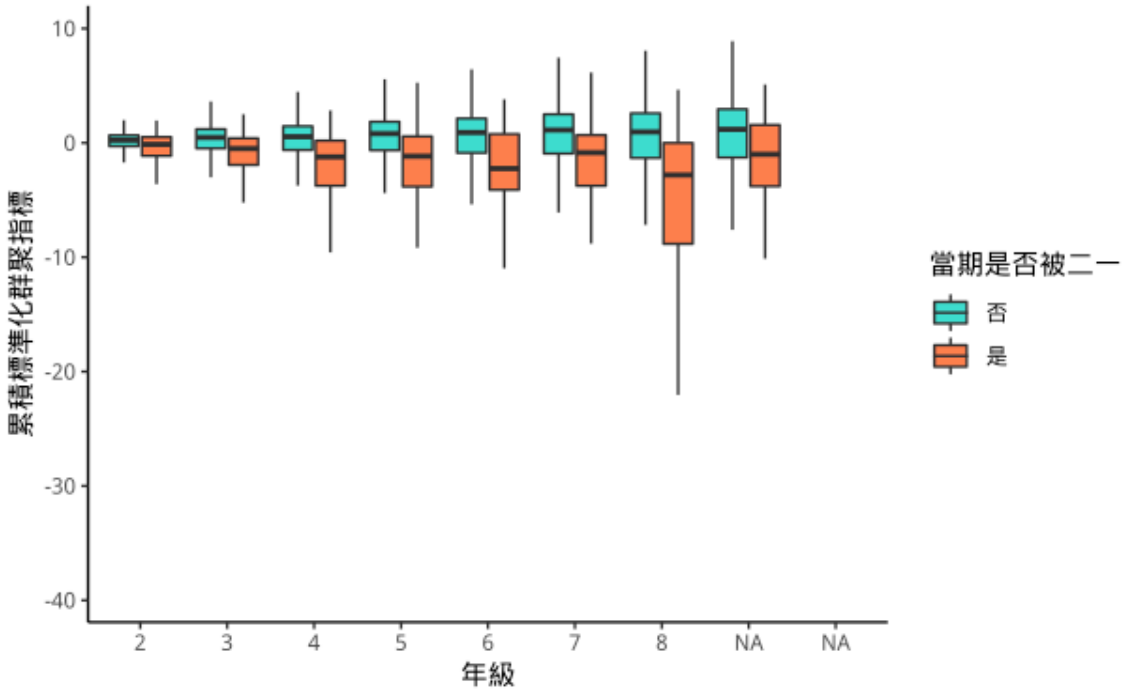


圖 2-11 顯示與我們推測一致，兩類型的學生在此特徵上有著明顯的差別，說明此特徵為一個良好的預測特徵。

第五項 綜合資料觀察

針對以上萃取出的變數進行相關性的觀察，並於此節對上述所有特徵整理說明。

表格 2-1：各特徵與是被預測變數相關係數表

	是否被二一
累積標準化群聚指標	-0.09
累積通識被當比	0.10
累積選修被當比	0.11
累積專業必修被當比	0.21
累積共同必修被當比	0.14
累積其他必修被當比	0.04
該學期選修修課比例	-0.08
累積專業排名	-0.13
累積其他排名	0
累積共同排名	-0.10
累積選修排名	-0.10
累積通識排名	-0.07
累積二一	0.16

由表格 2-1 顯示，可以清楚看到，累積排名特徵之間相互具有正相關，其中又以累積專業必修排名與累積選修排名兩者之間的相關程度最高，而兩者又分別代表本系專業知識的程度與在更深入專業知識中的程度，通常各系畢業學分中，對於外系選修的學分承認是具有上限的，故選修課程中大部分仍為本系專業課程，若一位學生於本系的專業科目中表現良好，那麼他在更深入的專業科目中也可以得到較好的成績。

各類累積排名特徵間的正相關以及各類累積被當比之間的正相關，反應了整體成績表現是會同上同下的，若一位學生對於大學的教育產生了不適應，那麼在整體

表現上皆會同時下降，故若能透過預測學生二一行為，儘早使校方採取預防措施，對於學校整體的教學是有利的。

綜合以上所有特徵的整理，首先我們觀察了過去被二一的次數，了解學生於未來時被二一的狀況，接著利用累積被當比來觀察一位學生於各類課程中的不適應狀況，然後更進一步利用累積排名特徵，了解學生程度差異的狀況，最後加入預測當期的課程繁重程度與學生心裡層面上的衡量指標，而所建的特徵中，兩類學生皆存在著差異性，皆為不錯的預測特徵。

第三章 研究方法

第一節 分類問題描述

本研究所面對的問題為：對於任意學生 i，給定他的特徵變數向量 x_i，要如何產生他當學期會不會二一的預測結果 $\hat{y}_i$，其中 $\hat{y}_i$ 為 1 或 0 值的分類變數，1 代表預測會被二一，0 則為否。令 $\mathbf{x}_i \in \mathbf{R}^d$，而 $\hat{y}_i \in \{0,1\}$，此分類問題為找尋一個函數映射關係

$$\mathcal{M}(\mathbf{x}_i \mid \theta)\colon \mathbf{x}_i \in \mathbf{R}^d \to \{0,1\},$$

估算其中的模型參數 θ 得到其估計值 $\hat{\theta}$ 後，並以此估計後之模型，以 $\hat{y}_i = \mathcal{M}(\mathbf{x}_i|\hat{\theta})$ 產生預測結果。

本研究使用四種函數模型進行預測估算，分別是：羅吉斯迴歸、隨機森林、支持向量機、人工神經網路。又隨機森林模型使用了決策樹樹型，故於以下小節我們針對這五種模型一一進行介紹。另外，模型在估算時只使用部份資料，稱之為訓練集，以 S_{train} 表示。

第二節 模型

第一項 羅吉斯迴歸 Logistic regression

令p_i為學生i於當學期會被二一的機率，即 $p_i = \Pr(Y = 1|X = x_i)$ 。羅吉斯迴歸假設此機率為如下函數形式：

$$\Pr(Y_i = 1|X_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)}$$

預測方式為：

$$\hat{Y}_i = \begin{cases} 1 & 若\ p_i \geq c \\ 0 & 若\ p_i < c. \end{cases}$$

c 為門檻值，若二一的機率高於此門檻值便預測該學生會為被二一，若低於此值則為預測不會被二一。對應到 $\mathcal{M}(x_i|\theta)$ 表示，此模型的參數為 $\theta = \{\beta, c\}$ ，其中參數 c 為此模型的超參數（hyperparameter）。

模型估計訓練方式

在給定超參數值 $\hat{c}$ 下，此模型估計方式為最大概似估計法，其求極值之目標函數為下

$$max_\beta \sum_{i \in S_{train}} \{y_i \log p_i + (1 - y_i)\log(1 - p_i)\}$$

$$\hat{p}_i = \frac{\exp(\hat{\beta}' x_i)}{1 + \exp(\hat{\beta}' x_i)}$$

對於超參數的估計我們一致放在第三節說明。

第二項 決策樹

預測方式為：

由特徵變數形成的一連串的樹狀決策結構，從根結點（root node）出發開始一連串由單一特徵變數所形成的「是/否」分枝，一直下去直到判斷出所屬分類。預測方式為依特徵值從根結點依行走，每遇到結點即依結點分枝的行走條件決定後續路徑，直到落在葉結點（leaf）為止。以圖 3-1 為例，此決策樹有一個根結點（累積專業必修被當比）及兩個葉結點（被二一、沒被二一）。若有位學生他的累積專業必修被當比為 0.45,，累積其他排名為 33；從根結點出發，由累積專業必修被當比進行第一次的分裂，其值低於 0.1 故會分裂至左側的葉節點， 最後預測為沒被二一。同樣學生若改以圖 3-2 例；此時位於左側的葉節點中不純度仍然太高，左側葉節點依累積其他排名再次特徵進行分裂，分裂至右葉節點，最後預測為被二一。

模型估計訓練方式

決策樹所需估算的問題為：要不要形成結點進一步分枝？如果要，結點要選什麼特徵值？分枝的條件是什麼？

以圖 3-1 及 3-2 進行說明。資料尚未經由特徵分裂前包含了所有的資料，即根結點，接著如圖 3-1 考慮的第一個分支特徵為累積專業必修被當比，經由特徵分類後會依此特徵值是否小於 0.1 此條件分裂出兩個葉節點。若分類過程於此處停住，並無繼續往下分裂，即僅進行一次分裂便得出最適分類，則其決策樹的深度為 1。若繼續往下分裂，會依照累積其他排名特徵再次進行分裂，並依照特徵值是否大於等於 40 分裂出兩個葉節點，如圖 3-2 所示，此時決策樹的深度則為 2。另外控制決策樹至多往下分裂次數的參數，稱為決策樹的最大深度，此參數為超參數。而主要需估計的參數為各結點所使用的特徵變數及分枝情境條件。

決策樹分類下，初始所有資料均位於根節點，然所有資料經決策樹分類後，個別會落在其中一個葉結點，故每個葉結點會搜集到一群資料。而完美的分類必需是每個葉結點資料群都是同類，即都被二一或都不被二一，若不是則有異類混雜的現象。因此，可進一步計算它每個葉節點群的異類混雜程度，即不純度的概念，接著再進一步去考慮要不要對某個葉節點改成特徵變數子結點，進一步分類，以降低整體的不純度，直到不純度夠低為止，訓練過程如下：

圖 3-1：決策樹第一次分裂圖

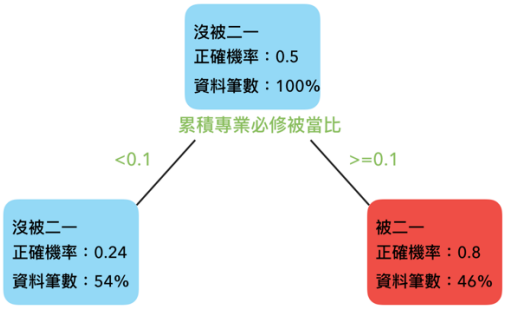


圖 3-1 中，右側葉節點的正確機率為 0.8，代表此葉節點的資料中，有八成的資料為被二一資料，模型經過不純度計算之後發現已夠低，便會停止往下分裂節點。左側之子節點裡，資料中沒被二一的比例僅有 0.24，故會繼續往下伸出節點，直到節點中的不純度夠低為止。

圖 3-2：決策樹第二次分裂圖

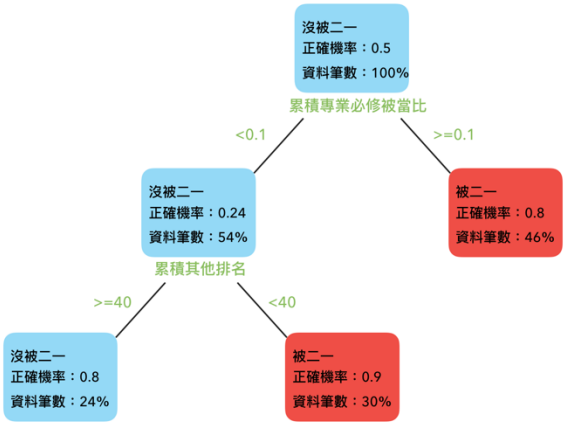


圖 3-2 中，左側子節點，繼續向下分裂，分裂出的兩個葉節，且不純度都以達到夠低，此時將停止分裂。

第三項 隨機森林

隨機森林由 Breiman Leo（2001）所提出，其基本原理為結合多棵決策樹，並加入隨機分配的訓練資料，以大幅增進最終的運算結果，此方法為 Ensemble Method（集成方法）的一類，其想法為如果單個分類器表現不錯，那麼將多個分類器組合起來，其表現會優於單個分類器，建構模型的步驟為:

1. 決定隨幾森林中需要樹木量 K：Hoerl A.E. and Kennard. R.W（1970）個數推薦介於 64 到 128。
2. 利用 Bagging 方式建造 K 棵決策樹：Bagging 於 1996 年由 Breiman 提出（Bootstrap aggregating），此種方法會從訓練集中取後放回隨機抽取 K 個樣本集，再從這 K 個樣本集中訓練出 K 棵數。
3. 由 K 棵決策樹共同預測應變數，出現最多的類別則預測的類別。

隨機森林的建構模型的方法是生成很多棵決策樹，由這些決策樹的結果去投票得出最終預測，其中這些決策樹必須有所差異，除了使用 Bagging 的方式讓 K 棵決策樹有所差異，在隨機森林的決策樹生成時，每棵樹在架構前可從總特徵變數中隨機抽取 q 個變數來當作架構此樹之分割變數的可選擇集合，造成樹之間的差異性。在隨機森林模型下， K 與 q 均為超參數。

第四項 支持向量機

模型設定：

考慮由特徵變數所形成的邊界函數 $f(x_i|\mathrm{w},\mathrm{b})$ ，預測方式為：

$$y_i = \begin{cases} 1 & \text{if } f(x_i|w,b) \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

此函數的分類預測完全依據學生的特徵變數落在邊界函數的哪一側決定 y_i 的類別。這裡 y_i 出象改成 {1,-1} 而非 {1,0} 是為了後續估計討論方便，並不影響預測結果。對應到 $\mathcal{M}(x_i|\theta)$ 此模型的參數為 $\theta = \{w, b\}$ 。

模型估計訓練方式

要了解支持向量機我們先以線性邊界函數開始說明，即

$$f(x_i\,|w, b) = w^T x_i + b,$$

此線性組合可形成一個超平面，給定 w、b 所決定的超平面我們可以計算所有資料特徵 x_i 與此平面的垂直距離，所有垂直距離中最小的即稱為此超平面與資料間的邊界差距（margin），此模型訓練過程可寫成：

$$\begin{aligned} &max_{w,b} && M \\ &s.t. && y_i(w^T x_i + b) \geq M \text{ for } i \in S_{train} \end{aligned}$$

若最適解下的 M 大於 0，則訓練資料可被特徵變數超平面完美區隔成 2 類。若訓練資料不可被特徵變數超平面完美區隔成 2 類，必需放寬邊界差距要求，允許一定的切割誤差，而成為如下的估計問題：

$$\begin{aligned} &\max_{w,b} && M \\ &s.t. && y_i(w^T x_i + b) > M(1 - \epsilon_i) \\ & && \epsilon_i \geq 0 \text{ for } i \in S_{train}, \sum_{j \in S_{train}} \epsilon_j \leq C \end{aligned}$$

C 為大於 0 的超參數， C 越大則模型對分類錯誤的容忍度也越大。

若超平面仍無法完整區隔成兩類，則有必要進一步引入「非線性」分類函數，使估計問題擴充如下：

$$\max_{w,b} \quad M$$

$$s.t. \quad y_i(w^T \sum_{j \in S_{train}} k(x_i, x_j|\gamma) + b) > M(1-\epsilon_i)$$

$$\epsilon_i \geq 0 \text{ for } i \in S_{train}, \sum_{j \in S_{train}} \epsilon_j \leq C$$

其中 k(·|γ) 為一非線性函數，稱之為核函數（kernel function），其參數 γ 也是超參數。

第五項人工神經網路 Artificial Neural Network (ANN)

模型設定：

ANN 為一種透雙層函數映射來捉捕預測函數的非線性特質，其將預測模型函數設定成：

$$\mathcal{M}(\mathbf{x}_i \mid \theta) = f(g(\mathbf{x}_i))，$$

進一步可再表示成：

$$x_i(K \times 1) \xrightarrow{g} n_i(M \times 1)$$

$$n_i(M \times 1) \xrightarrow{f} p_i(1 \times 1)$$

其中 g 函數所映出的值稱之為神經元，更詳盡的定義如下：

$$g(x_i) \equiv \begin{bmatrix} g_1(x_i) \\ g_2(x_i) \\ \vdots \\ g_M(x_i) \end{bmatrix} = \begin{bmatrix} a(w_1{}'x_i + b_1) \\ a(w_2{}'x_i + b_2) \\ \vdots \\ a(w_M{}'x_i + b_M) \end{bmatrix} \equiv \begin{bmatrix} n_{1,i} \\ n_{2,i} \\ \vdots \\ n_{M,i} \end{bmatrix} = n_i$$

函數 a(z) 一般稱為激活函數，用來控制 z 是否輸出，常用的函數型態為 $a(z) = max(0, z)$ ； 此外，我們選定 $f(z)$ 為一 logistic 函數:

$$f(n_i) = \frac{\exp(\beta' n_i)}{1 + \exp(\beta' n_i)}$$

預測方式為：

$$\hat{Y}_i = \begin{cases} 1 & 若\ f(n_i) \geq c \\ 0 & 若\ f(n_i) < c. \end{cases}$$

c 為門檻值，若二一的機率高於此門檻值便預測該學生為會被二一，若低於此值則為預測不會被二一。對應到 $\mathcal{M}(x_i|\theta)$ 表示，此模型的參數為 $\theta = \{w, b, \beta\}$ ，超參數（hyperparameter）為學習率，較大的學習速率，有較大的網路加權值修正量，可較快逼近函數最小值，但過大的學習速率將導致網路加權值修正過量，造成較難達到收斂。

模型估計訓練方式

在給定超參數下，此模型估計方式為最大概似估計法，其求極值之目標函數為下

$$max_{\mathbf{w},\mathbf{b},\beta} \sum_{i \in S_{train}} \{y_i \log p_i + (1 - y_i)\log(1 - p_i)\}$$

第三節 模型訓練

本文將資料分割成兩個資料集，訓練集以及測試集分別佔總資料 70%與 30%，測試集僅於最後選定模型之後套入模型使用，訓練集主要為訓練分類器；並透過 10 疊交叉驗證集來進行超參數的訓練。

參數與超參數的區別為，參數是指選定的機器學習技術中用來調整資料的變數，如本文所使用支持向量機中的 w、b ，而超參數用在避免參數選擇形成過度配適現象，即估計模型在訓練集過度、虛假的表現良好，但在測試集有顯著預測失準

現象。超參數的選擇，如本文所使用支持向量機中的 C、γ ，乃利用驗證集的預測準確調校達成。

第四節 不平衡資料調整

不平衡資料指的是資料中類別的不平均，以本文訓練集資料為例，資料中含有 7978 位學生，而被二一的學生類別僅有 618 位，兩者比例相差懸殊為典型的不平衡資料。若是分類模型是以總體分類的準確度為目標進行學習，將導致預測重點過度著重於多數類資料分類（即這裡沒二一類別）的準確度，從而使得少數類樣本的分類能力下降，故絕大多數常見的機器學習演算法對於不平衡資料集都不能很好的分類。

為資料類別不平衡的問題，本文於訓練模型前，使用多數樣本不足抽樣（Majority Under-sampling）及少數樣本過度抽樣合成法（Synthetic Minority Over-sampling Technique），簡稱為 SMOTE[2] 來達成平衡樣本；前者由沒被二一類資料點採取隨機抽樣，但抽出樣本數少於原樣本數，後者針對被二一的少數樣本進行樣本增加，直至兩類資料筆數相同為止，最後被二一樣本與沒被二一樣本皆有 1614 筆。

第五節 模型預測表現衡量

目前對於演算法模型評價的指標又很多如：召回率、準確度、F 值、Area Under Curve （AUC）、R square 等，多數應用於教育資料探勘文獻中的指標為準確度與 AUC ，其中 AUC 指得是 ROC 曲線（receiver operating characteristic curve）下方的面積。

早期 Receiver operating characteristic curve （ROC 曲線）主要利用於生物醫學上，後也開始被廣泛使用於機器學習，在相關的驗證研究的文獻當中， ROC 曲

[2] Chawla,Bowyer,Hall and Kegelmeyer (2002)。

線可謂最常被使用來驗證整體模型效度之方法，透過調整其閥值來衡量分類正確與錯誤的次數，藉此來呈現二一學生捕捉率（在文獻中稱為敏感度），與誤查率之間的抵換關係。以下將 True Positive、True Negative、False Positive、False Negative 簡稱為 TP、TN、FP、FN。

表格 3-2：混淆矩陣

實際 / 預測	二一	沒二一
二一	TP	FP
沒二一	FN	TN

ROC 曲線之橫軸為誤查率，縱軸為二一學生捕捉率，二一學生捕捉率定義為 $\frac{TP}{TP+FN}$ ，所刻畫的是分類器所分類出的被二一占實際上被二一的比例，誤查率定義為 $\frac{FP}{FP+TN}$ ，刻劃的是分類器分類出沒被二一的學生卻被誤人為被二一，占實際為沒被二一生的比例。 此兩指標具有抵換關係，假設原模型被視為正例的閥值為 0.5，即模型預測該點為正例的機率大於 0.5 便為正例，反之則為負例。如果減少閥值至 0.1，則能識別出更多正例，也就是提高了二一學生捕捉率，但同時也會將更多的負例預測為正例，即降低了誤查率，在統計學上又將 FP 稱為「型一錯誤」，FN 稱為「型二錯誤」。 模型在訓練各種不同的閥值時，會分別有各自的一組二一學生捕捉率與誤查率，將其各點的二一學生捕捉率與誤查率繪於圖上行形成 ROC 曲線，隨著閥值的遞增，二一學生捕捉率和所對應的誤查率均會呈現遞增狀況。

圖 3-3：型一型二錯誤關係圖

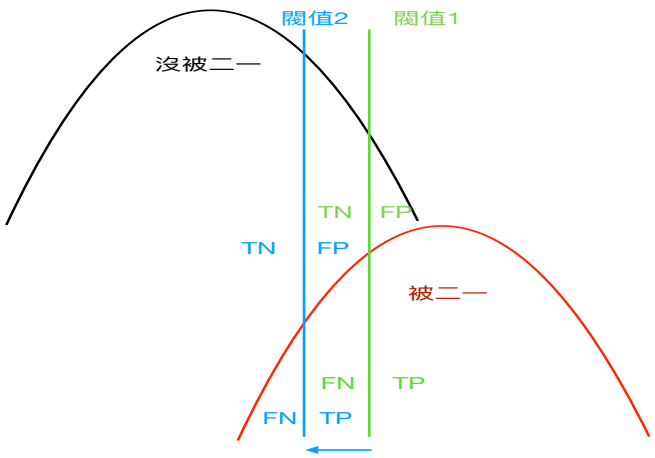
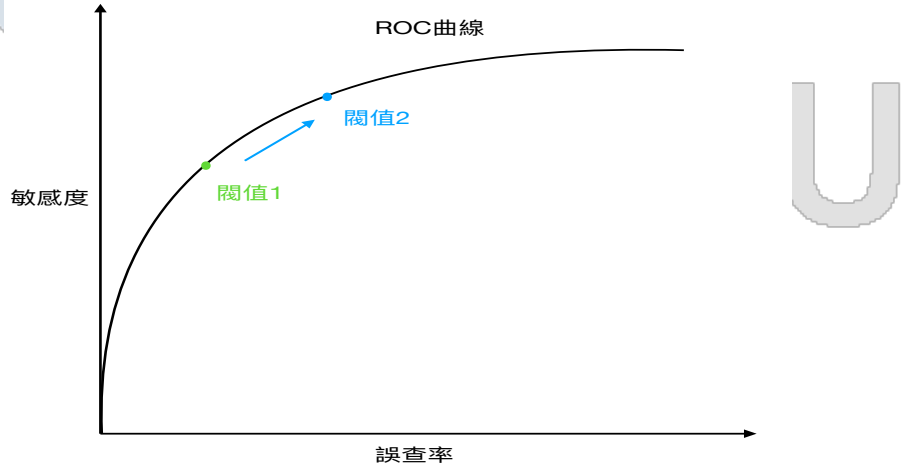


圖 3-4：ROC 關係圖



ROC 曲線提供了一個方便觀察模型優劣的方法，首先觀察 ROC 曲線圖的幾個點；若為（0,0）則代表該分類器會預測所有樣本皆為負例，（1,1）則是將其全數預測為正例，（0,1）代表的是將所有的樣本都正確分類，（1,0）則為全樣本錯誤分類，可以得出 ROC 曲線若能向左上角靠則代表此分類器擁有越好的分類效果，當一個模型的預測能力完美時， ROC 曲線會在左方與縱軸貼齊，上方與橫軸貼齊； ROC 曲線越往左上則代表分類效果越好，而當 ROC 曲線貼合於從原點出發的對角線時，則代表此模型為隨機模型，即此模型沒有任何預測能力，而當 ROC 曲線之間出現交錯或難以觀察的狀況時，可藉由 ROC 曲線下的面積 AUC（Area Under Curve）來評定分類器的好壞，即 ROC 曲線下的面積占整體橫縱軸所構成之四方形面積的比例，若為 1 代表完美模型，0.5 則代表此模型毫無預測能力，一般模型此值皆介於 0.5 至 1 之間，若於 0.5 以下則代表該分類器具有相反的分類效果。

準確度以及 AUC 評價指標之間的抉擇， Bradle 於 1997 年文獻中進行了比較；文中表示我們應優先選擇 AUC 應用於機器學習中預測結果的評價，主要可以依據下列理由，準確度會因為不同的閥值（如羅吉斯迴歸所使用的 c 值）而導致有不同的結果而 AUC 是各閥值所連起的線下面積；故 AUC 不會受到閥值選擇的影響。在陽性資料（即被二一）與陰性資料（即沒被二一）差距很大時 AUC 擁有著比準確度還好的評價功能，例如陰性類沒被二一的樣本佔大多數時，模型預測力可輕易以高閥值使預測結果多為陰性而輕易提高。故本文分類器驗證指標將使利用 AUC 來作為評價標準。

準確度公式如下：

$$\frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

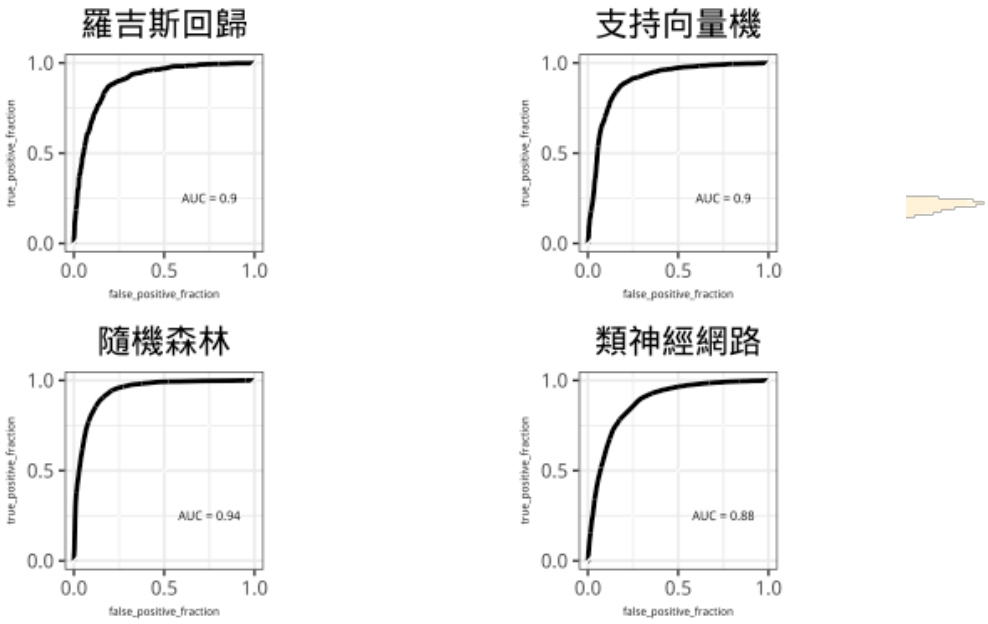
然而當資料為不平衡資料；如本文資料中沒被二一類的資料遠高於被二一類，即 Positive 類較少，即使分類器將全部的資料都遇測為沒被二一類（Negative），

其準確度仍可以達到很高，故本文分類器驗證指標將使利用 AUC 來作為評價標準。

第四章 研究結果

第一節 模型結果

圖 4-1：模型 AUC 比較圖



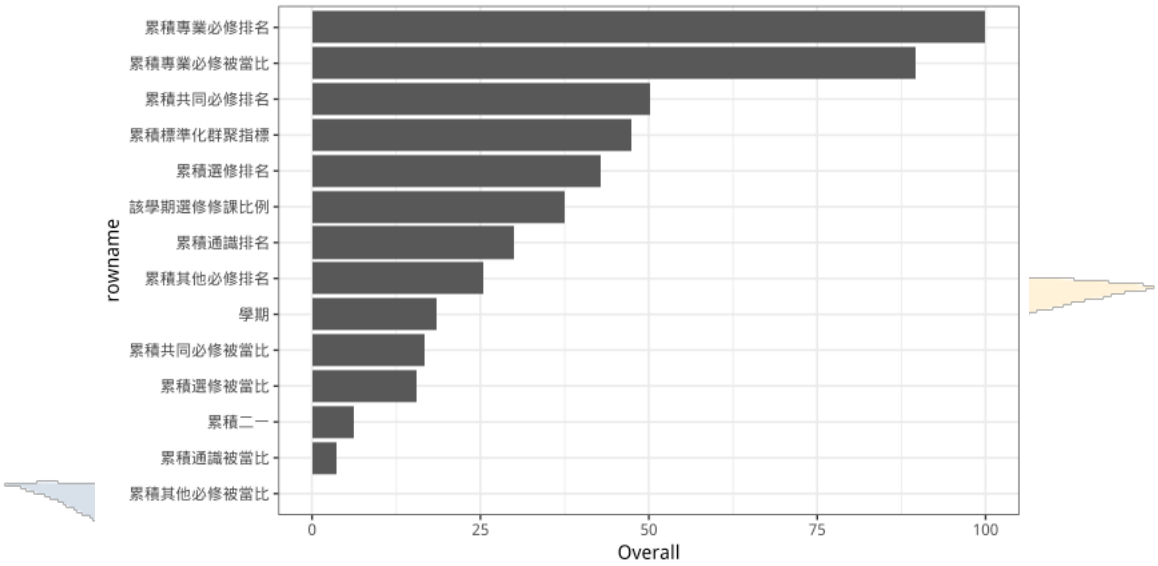
分別比較羅吉思迴歸、支持向量機、隨機森林以及類神經網路之 AUC，圖 4-1 中顯示，四個分類器的分類效果表現皆有 0.8 以上，代表預測的特徵選取不錯，其中 AUC 最高者為隨機森林模型，值為 0.94，故將使用隨機森林於測試集資料中進行測試，再於第三小節中與基礎模型進行比較。

第二節 變數重要性衡量

得知隨機森林為以上幾個中最好的分類器後，利用特徵經過置換前與置換後的誤差影響，來衡量各特徵的重要性。 結果如圖 4-2 所示，觀察解釋能力前三重要的變數為：累積專業必修排名、累積專業必修被當比、累積共同必修排名。由於專業必修與選修所代表的是本系的專業知識，說明專業性的知識對於學生是否會被二

一的影響程度是最大的，除了本系專業科目至關重要外，圖 4-2 也反應了，全校共同必修課的學習狀況對於學生日後是否會被二一也有重要的影響，若校方可以針對學生的共同必修科目以及專業必修科目加強，相信能大幅降低學生被二一的人數。

圖 4-2：變數重要性圖



第三節 基礎模型比較

由上小節得知，最重要的變數為累積專業必修排名。本文僅用變數累積專業必修排名，製作羅吉斯迴歸作為基礎模型，並以混淆矩陣在測試集的表現與隨機森林進行比較，結果如表格 4-1。

隨機森林的誤查率僅有 0.1562，平衡準確率（Balanced Accuracy[3]）為 0.8251，而基礎模型的誤查率以及平衡準確率分別為 0.2749 與 0.7845，隨機森林於此兩指標的表現皆比基礎模型來得好，雖然隨機森林的二一學生捕捉率 0.8064，低於基礎模型的 0.8440，但是在不犧牲太多的二一學生捕捉人數下，可以大幅的提升分辨沒被二一類的學生是可以接受的。綜合以上本模型在於預測學生被二一有著很好的預測效果。

[3] 即 $\frac{\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)}{2}$

表格 4-1：混淆矩陣比較

	羅吉斯迴歸（單一變數）		隨機森林	
實際 / 預測	二一	沒二一	二一	沒二一
二一	157	2885	150	1639
沒二一	29	7607	36	8853

第五章 結論與建議

本文分別利用了四種分類器，分別為羅吉斯迴歸、隨機森林、支持向量機以及人工神經網路，預測學生未來的二一狀況，結果顯示隨機森林在預測中表現較好；平衡準確度為 0.8251，召回率為 0.8064，本文貢獻在於透過隨機森林模型可以即早的找出未來被二一可能性較大的學生，並且透過觀察變數間的重要性發現本科專業知識的學習狀況對於一位學生未來是否會被二一有著很大的關係。校方未來可以運用本文模型進行預測找出未來會被二一的學生，並且規劃出一個完善的補救措施來提升學生專業知識的表現，相信可以達到降低學生未來被二一的比例，提升學校品質並減少退學率。

本文所利用成績單資料預測學生於未來時是否會被二一之表現雖得到不錯的效果，但許多時候學生是否會被二一也會由學生之心理狀況、學生出生背景；如父母職業，是否為明星高中，出生地人口密集程度，以及學生大學前成績表現等狀況影響，若未來想增強預測能力除了增加樣本資料外，亦可藉由此方向著手。

第六章 參考文獻

中文文獻：

吳東陽（2018）。大學雙二一退學與學生行為。私立東吳大學經濟系研究所碩士論文。

鄭媛文（2013）。同儕教導學習策略對學生學習成就與情意態度影響之後設分析。教育理論與實踐學刊第 28 期 。

陳家琪（2017）。從期中預警與學期成績之關係談預警成效—以臺北市立大學為例。臺北市立大學數學系數學教育碩士論文。

李勁昇（2017）。探討學生的選課身分別、修課習慣與成績之關聯性－以臺北市立大學為例。臺北市立大學數學系數學教育碩士論文。

李易倫（2015）。大學生學業成績對同儕人際關係與課堂焦慮的影響：以正向心理資本為調節。國立東華大學國際企業學系碩士論文。

梁玲（2016）。學生學習動機與同儕學習對學習成就的影響。高苑科技大學經營管理研究所碩士論文。

英文文獻：

Ghadeer S. Abu-Oda and Alaa M. El-Halees (2015). Data Mining In Higher Education: University Student Dropout Case Study. International Journal of Data Mining & Knowledge Management Process (IJDKP) vol. 5, No.1, January.

P. Baepler and C. J. Murdoch (2010). Academic Analytics and Data Mining in Higher Education. International Journal for the Scholarship of Teaching and Learning, vol. 4, no. 2, pp. 1-9.

S. Kotsiantis, C. Pierrakeas and P. Pintelas (2004). Predicting Students Performance In Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence, 18:411-426.

Schaffer, C. (1994). A conservation law for generalization performance. In Proceedings of the Eleventh International Conference on Machine Learning, Pages 153-178, New Brunswick, USA, July 10-13.

Cortez, P. , and Silva, A. , (2008). Using data mining to predict secondary school student performanc.

Gerben W. Dekker, Mykola Pechenizkiy and Jan M. Vleeshouwers (2009). Predicting Students Drop Out: A Case Study .International Conference on Educational Data Mining (EDM) , 2nd, Cordoba, Spain, Jul 1-3.

B. Baradwaj and S. Pal (2012). Mining educational data to analyze student' s performance. Internation Journal od Advamced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

Breiman (2001), L., Random Forests. Machine learning, 45(1), 5-32.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321 – 357.

Andrew P Bradle (1997). Pattern Recognition, 30(7), pp. 1145-1159.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321 – 357.

第七章附錄

變數重要性衡量算法：

1. 利用每棵樹的分類模型來預測自己的 Out-Of-Bag（OOB）樣本，並計算錯誤率。

   OOB：在建構每棵樹的時候，我們對訓練集使用了不同的 bootstrap sample。所以對於每棵樹而言，大約有 1/3 的資料點是沒有參與該棵樹的生成，他們就是該棵樹的 OOB 样本。

2. 對想了解該特徵重要性的特徵進行隨機打亂。
3. 利用原隨機森林模型進行預測得到新的預測值。
4. 計算每棵樹新的 OOB 樣本錯誤率。
5. 對於每棵樹擾亂特徵前後所得到的錯誤率相減並平均。
6. 得出因該特徵擾亂後而導致的平均誤差上升多少，越高代表該變數越重要。