國立臺北大學經濟系

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

研究生：李冠緻

中華民國 108 年 7 月

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：36

所組別： 經濟 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，將資料探勘應於與教育之中，不僅可以儘早發現學生的問題並給予協助，也可以知道何種因素是影響學生日後被二一的關鍵。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%。經過變數的重要性衡量，得知決定學生未來是否會被二一，最主要的關鍵為累積專業排名。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狀況的影響最為嚴重，故校方可以針對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。

英文摘要：

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的措施，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普遍為「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才會退學。

交通大學註冊組組長彭淑嬌指出，十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，校方只好又改採雙二一制度，雖近期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) 研究也指出二一制度仍然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，依序觀察與介紹。

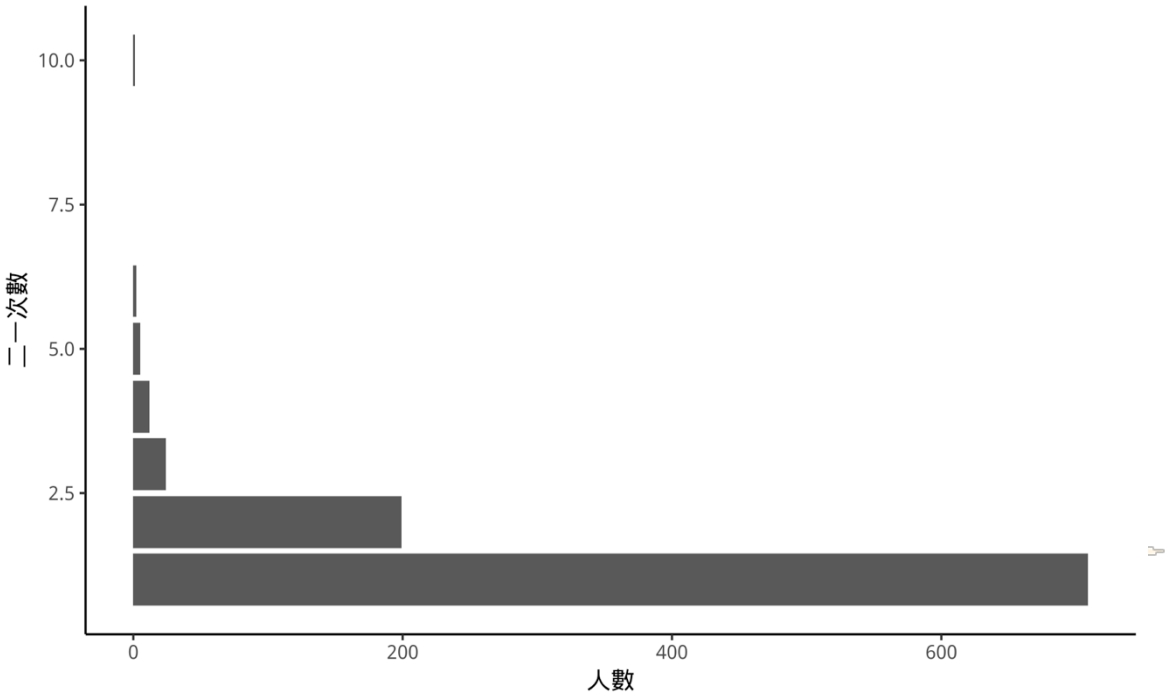
第一項 預測變數二一狀況觀察

二一頻率觀察

觀察校內被二一的學生狀況為何，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，被二一的學生人數有 952 位，佔學生人數的 8.8%，兩類預測值資料筆數相差甚大，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。

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，若資料中出現學生被二一次數超過兩次則代表該位學生為僑生或是身心障礙學生，由圖 2-1 可知出大部分的學生在被二一壹次後，會設法減少自己再次被二一的可能；二一制度對學生具有一定的警惕作用。

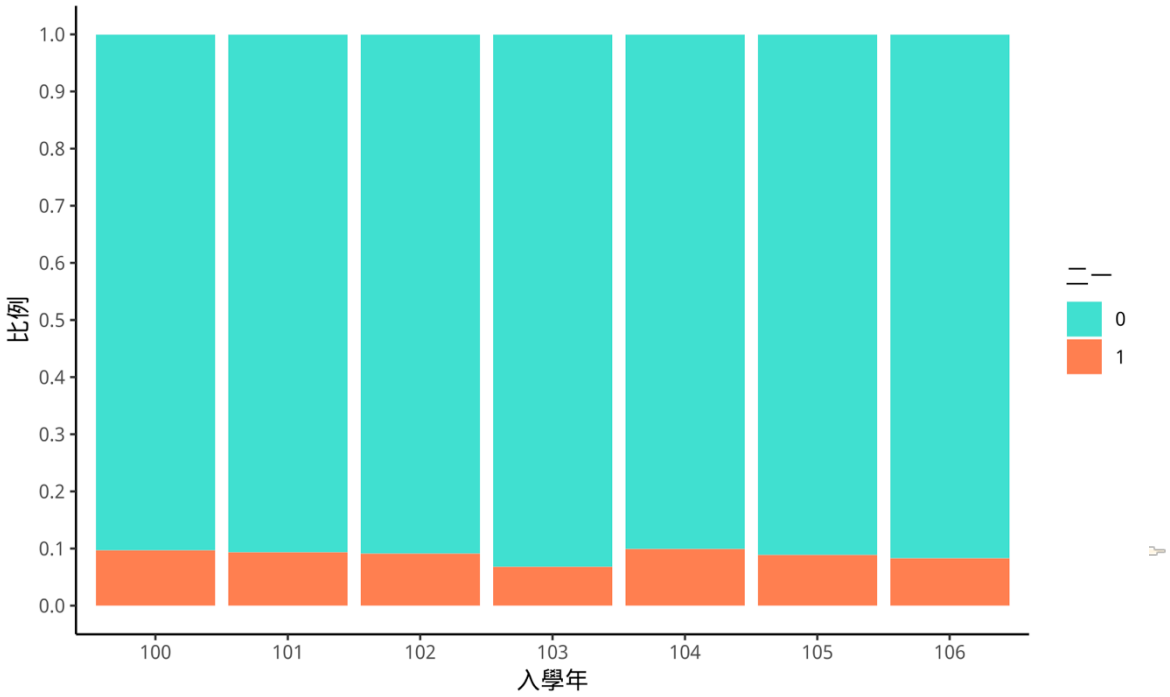
圖 2-1：學生二一頻率圖



二一時間性觀察

考慮學生是否會因為不同入學年此類的時間性因素，導致不同入學年學生被二一的比例有所差異，圖 2-2 顯示了入學年度與被二一的關係，可知各入學年在二一比例上相對隱定，皆約佔一成，代表著入學年是學生被二一 的因素的機會不高；此反應了每個入學年中皆有一部分的學生對於大學的教育出現了不適應的狀況。

圖 2-2：入學年與學生二一關係圖



因原始成績單的變數較為複雜，我們自原始成績單中各欄位改建為較為有用的特徵，以前幾期的特徵預測未來學期同學是否會被二一，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，故預測特徵可分為預測學期以前的訊息以及預測學期初的訊息，下小節將分別介紹各類特徵。

第二項 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

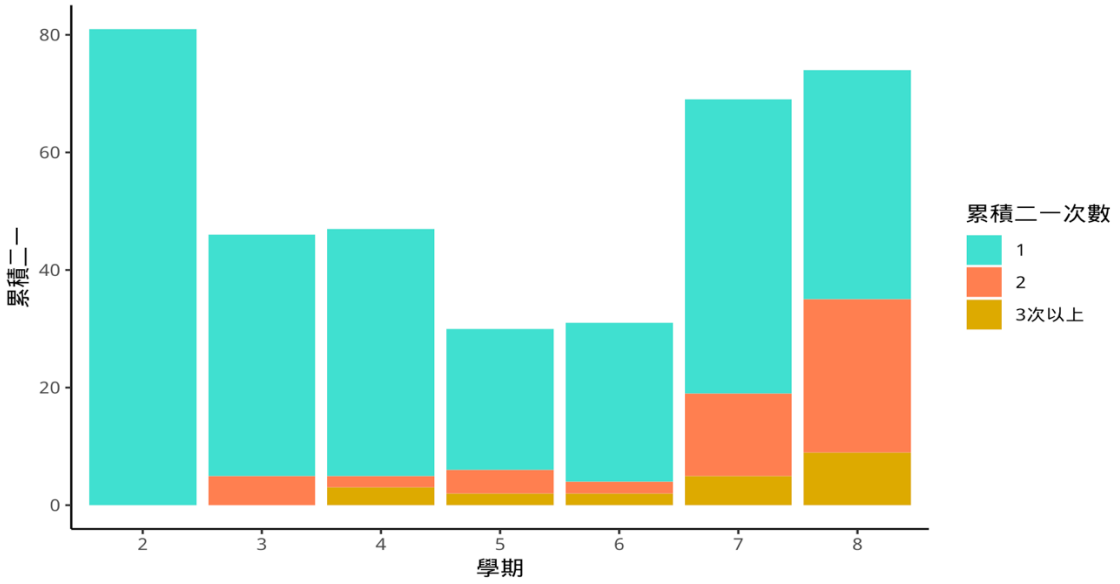
圖 2-3 為被二一的學生中，累積至預測當期前的二一狀況，橫軸數字代表第幾學期，圖顯示，學生於一年級時發生被二一的狀況是最多的，其反映了被二一的學

生中，有很大的比例對於初上大學的教學與生活是很不適應的，而被二一人數另一處高峰位於大四，在大四時被二一的學生分為兩種：

- 第一種為，大四修課數較少而導致，修課數較少狀況下，若有幾門課被當便很容易造成被二一；為圖中橘色的部分。
- 第二種為橘色以外的部分，這些學生都是曾經有被二一紀錄的學生，而到大四時又再度被二一；由此可知被二一過的學生大四時有很大的機率會延續之前的學習狀態導致大四再度被二一。

為了捕捉到這些大四而又再次被二一的人，我們將加入變數「累積二一」當作解釋變數，以預測同學未來是否會被二一。

圖 2-3：學生累積二一與學期圖



本節中指出了，有些學生的學習模式是會延續下去的，尤其在大四時，發生被二一的狀況最為嚴重，而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以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表現。

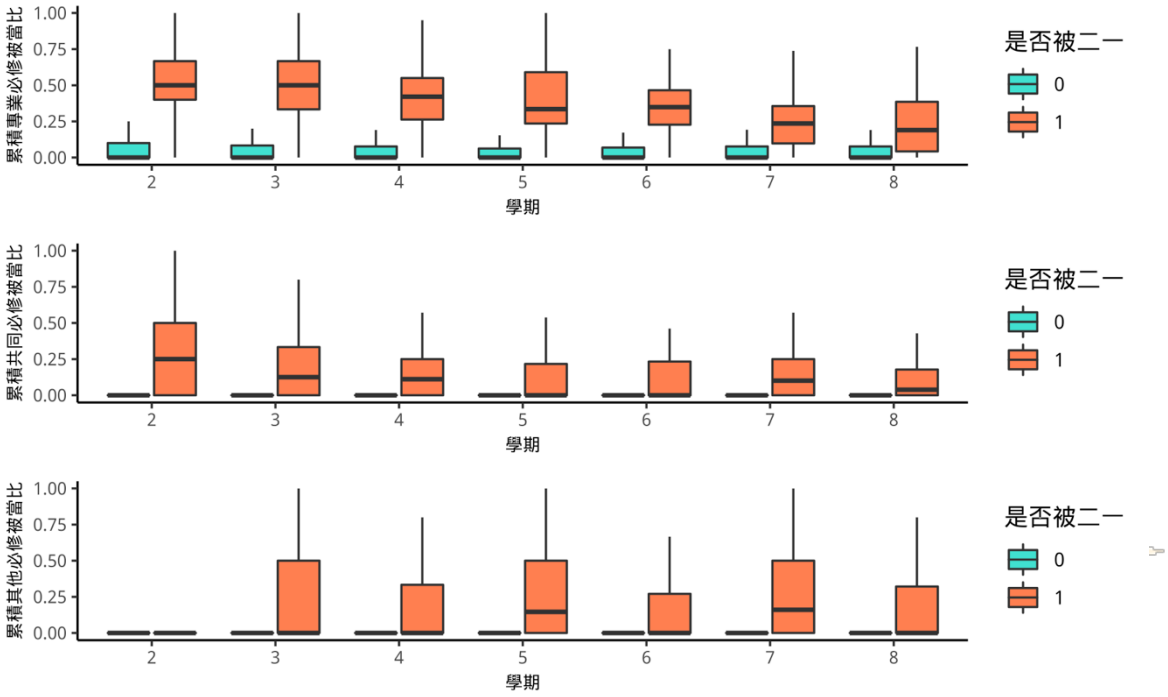
累積被當比

在課程類型上，大類可分成「必修」、「選修」及「通識」，其中必修又可細分成專業必修、共同必修、其他必修三類。 其中專業必修[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。

而在這五類課程我們去統計學生到前學期為止的適應狀況，所使用的適應狀況特徵為，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，每人的修課數會受到系級的不同、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。

[1]由於全校學系眾多，各系必修科目又有可能調動，故在判定上是以該屆同學有超過一半以上修習之必修科目名稱，即定義為該生所屬學系之專業必修科目；舉例來說：100 學年入學之經濟系學生有超過 5 成，其成績單上有出現標示為「必」修的「個體經濟學」成績，則「個體經濟學」會判定為此學年入學經濟系學生之專業必修。

圖 2-4：累積必修被當比與學生二一關係圖

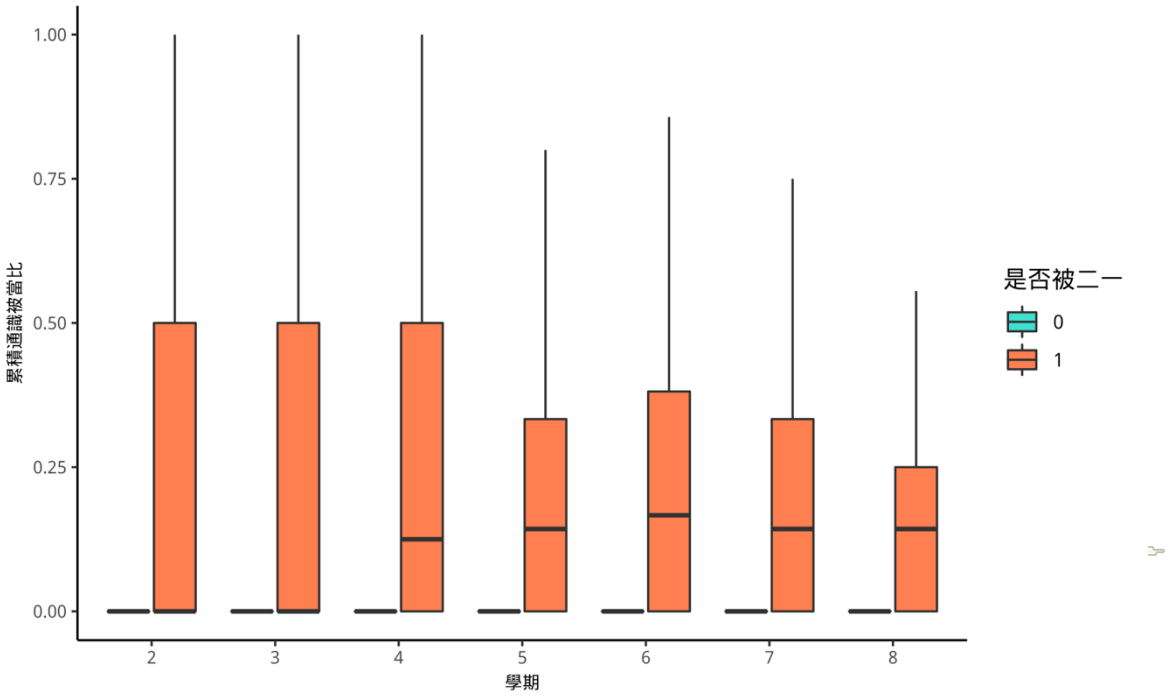


累積專業必修被當比可以看出一位學生在本系專業領域的修課狀況，由圖 2-4 可知於專業領域上，沒被二一的學生表現明顯比被二一者好，而也因專業領域上的知識通常較深所以會導致專業被當比無法像其他兩種類型的必修課一樣，其被當比的第 75 百分位數為 0。由此可知一位學生若在以往的專業科目上表現良好，那麼可以預期到，他在未來被二一的可能性是較低的。

於累積其他必修類別中，大一至大二時兩類學生的被當比第 75 百分位數皆為 0，其主要原因為大部分學生在大學初期較不會申請如教育學程之類的外系專業課程，而在大三至大四時，此特徵可以捕捉到，學生於外系專業上的不適應會導致未來被二一可能性提高。

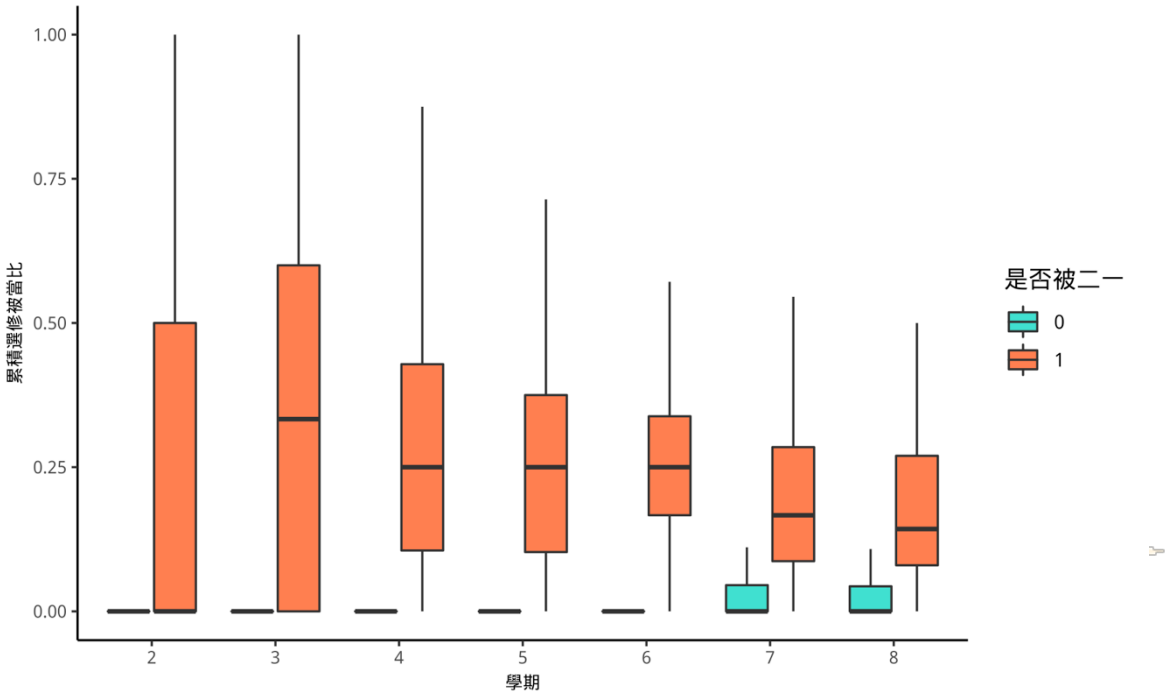
必修類裡最後一類為累積共同必修被當比，所捕捉的是同學對於學校共同科目被當的狀況，而全校共同科目必須顧及全校各系的學生程度，所以通常會是一門相對不容易被當的科目，若此類型的科目被當給出的訊號為學生對於大學生活的不適應，包括時間控管、等等非學習上的因素，圖 2-4 中清楚呈現，沒被二一類型的學生被當比第 75 百分位數皆為 0，代表此類裡大部分的學生，對於大學生活適應程度不算太糟，而被二一類學生被當比的第 75 百分位數高於 0，則顯示了，以往對於大學的不適應狀況越與未來被二一的擁有正向的關係。

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，從圖 2-5 顯示未來容易被二一的學生，對於以往的課外探索，相對於沒被二一的學生，較不重視，沒被二一類學生此比例與被二一類學生的中位數差異不大，主因為此類課程的難易度較低，通常同學於此類課程較不易被當，除此之外，通識科目也給了學生許多選課彈性，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 中顯示了兩類學生於此特徵的表現，可以看到兩類學生的第 75 百分位數差異很大，選修課程也給了學生一部分的選課彈性，故若學生選修被當比很高，我們傾向相信學生於未來被二一有較高的可能性。

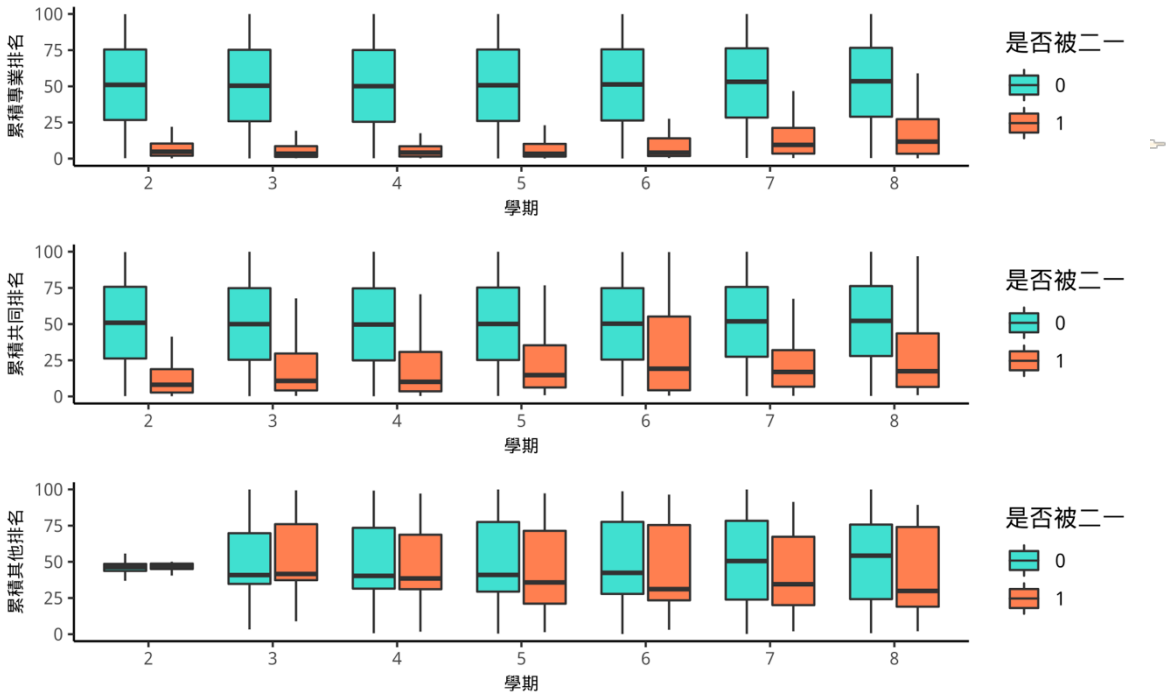
綜合上述分析可知，兩類學生在累積被當比變數群中的差異顯著，說明了它們是合理的預測二一特徵變數，了解修課狀況後，下小節將進一步從各類課程中討論成績所帶來的影響。

成績表現

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，故必須藉由觀察各類課程上的排名，才能了解修課上的真實狀況，觀察以往的成績排名對於未來是否

被二一影響為何？ 在此與上小節一樣針對不同的課程類別進行分別討論。成績計算方式如下：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，以總成績與同班的同學進行排名，同班的定義為，同系、同年級且同入學年；將成績的高低排名，映射至 0 到 100 區間，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

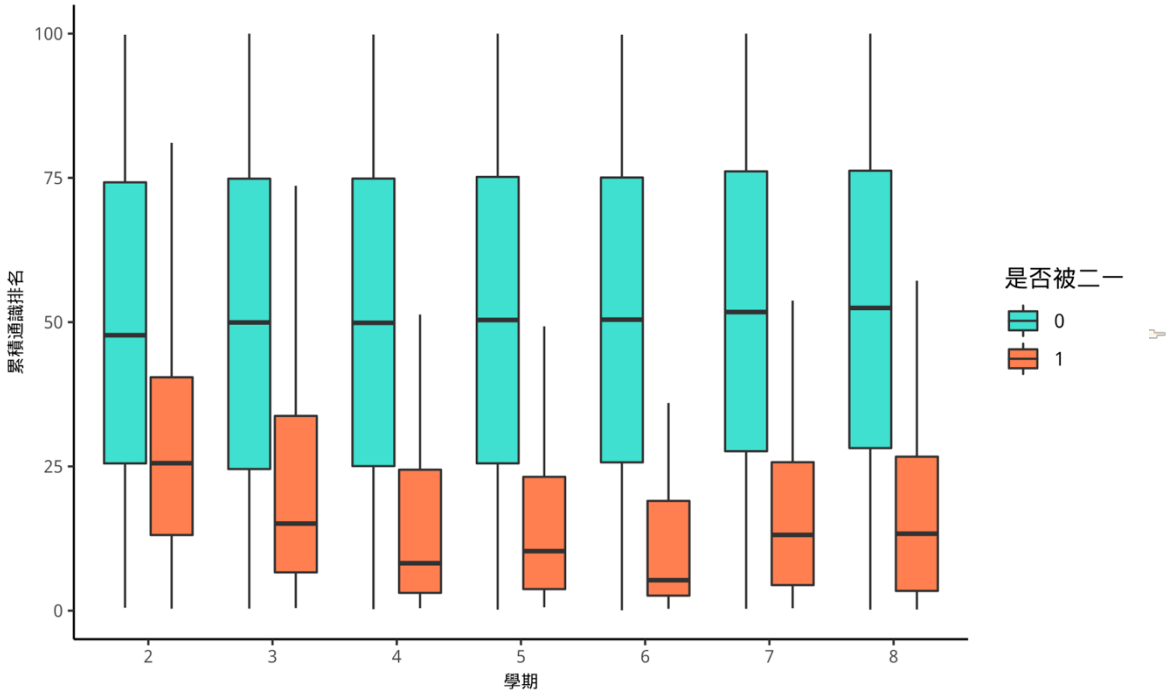


累積專業必修排名所代表的意義為，對於學生的專業知識學習狀況做排序，由圖 2-7 觀察，被二一的同學於以往的專業知識表現上低於沒被二一生許多；第 75 百分位數皆低於 25%，而在大四下時，被二一類學生的累積專業必修排名中位數顯著高於前幾期，顯示部分被二一生在大四上時，有嘗試的想讓自己的專業知識提高，但由於學習上的力不從心，導致雖有提升但是至大四下時仍被二一的狀況。

累積共同必修排名，代表著學生在學適應狀況的排序，圖 2-7 顯示，被二一的學生於以往的在學適應狀況會低於沒被二一生與前小節相互呼應。

累積其他必修排名較為特別，修此類別課程的學生人數相對其他兩類較少，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是無法被觀察到的。

圖 2-9：累積選修排名與學生二一關係圖

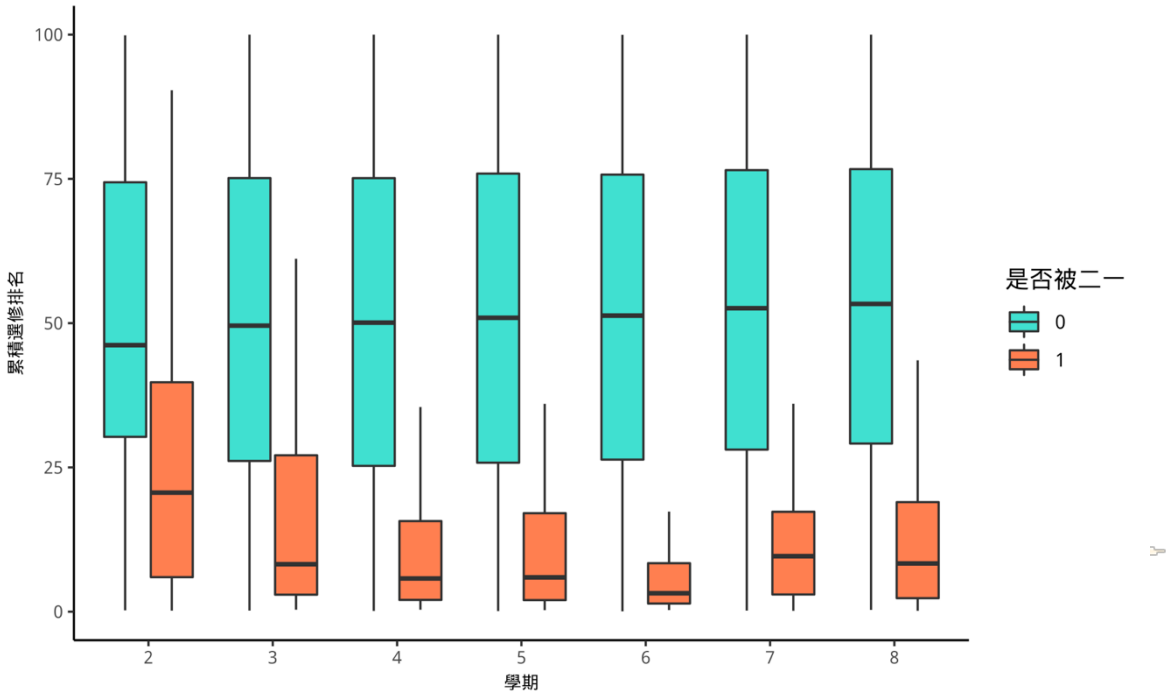


圖 2-9 為累積選修排名，兩類學生差異很大，可以作為一個很好的預測變數。

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，經過排序可以進一步劃分出，學生的排名狀況，藉由排名狀況的差異可以補捉到兩類學生的差異性，故將此類成績衡量變數作為一預測變數是合適的。下小節進入預測學期當學期可以提供何種訊息，即站在學期初我們要預測一位學生，除了前幾期特徵外，還有哪些特徵是可以得到的。

第三項 預測學期當學期訊息

研究目的為預測當學期結束前預測該生二一狀況，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，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和，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，該位學生班上同學於畢業時，所選修之選修課數的中位數；以此中位數代替畢業時選修規定堂數的真實值。

圖 2-10：修課比例與學生二一關係圖

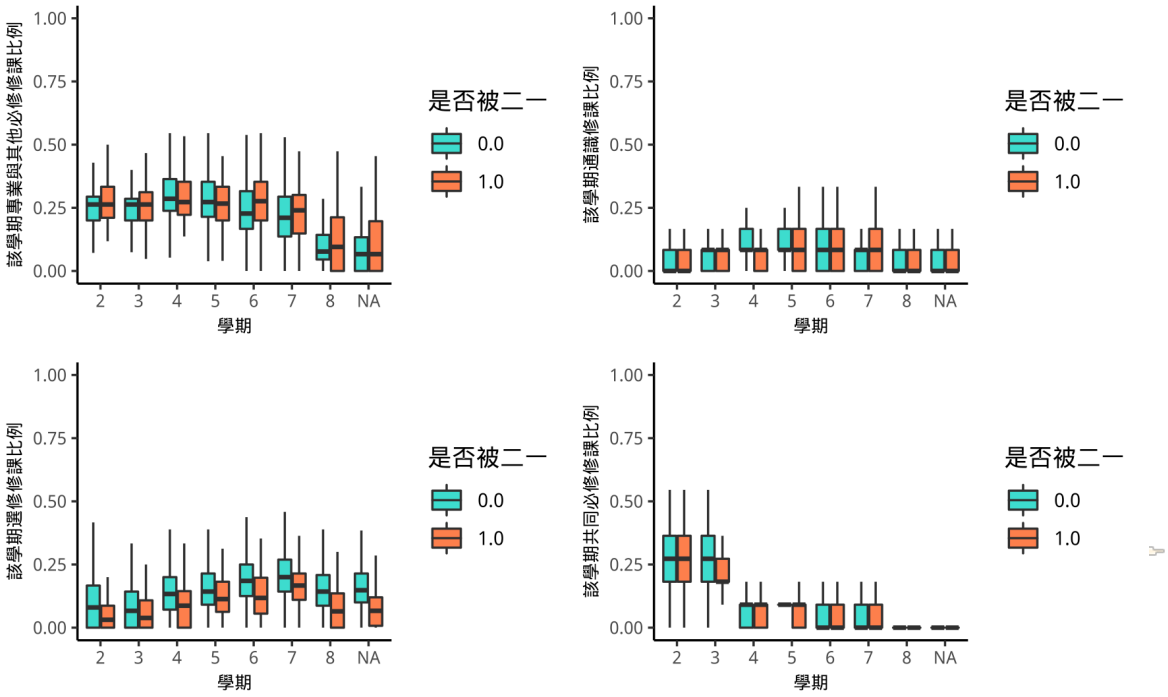


圖 2-10 中，除了該學期選修修課比例可以看出中位數擁有較明顯的差異外，其餘變數相對來說看不出趨勢為何，故推論，修課的繁重程度對於學生的影響沒有太明顯，作為一個好的預測變數，必須能夠在兩類學生中清楚得看出差異，而該期選修修課比例在此方面表現良好，故僅放入選修修課比例作為預測變數。

綜合以上皆在討論成績與修課方面的特徵，而研究顯示，同儕對於學生學習也存有很大的關係，下小節將討論同儕面向所帶來的影響。

群聚指標解釋變數

被二一的學生除了成績上的原因以外，也很有可能是心理層面上的問題，部分被二一的學生會想逃避同儕間所帶來的壓力，有意的安排，使自己不與班上同學修相同的課程，亦或是此類學生於班上的活躍度較低，與班上同學脫節，時常一人修課，而非跟著團體選擇修課，導致他所修的課程與大部分的同學皆不同。而此心理層面的資料較為敏感，也取得不易，故藉由以下所建立的特徵，試圖捕捉學生心理層面的狀態。

群聚指標是一個反應某位同學與班上同學間交集頻率的指標；其定義為，於學期結束後，所修的課裡分別有幾位班上同學一同修課，將其依依加總得到此指標。

累積群聚指標與前面小節的累積特徵概念較為不同，因於學期初時可以清楚的知道哪些同學有修相同的課程，故可計算當期的群聚指標，累積群聚指標是指累積至預測當期（包含預測當期）的群聚指標，群聚指標會受到各班各系修課數、畢業學分影響；在此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

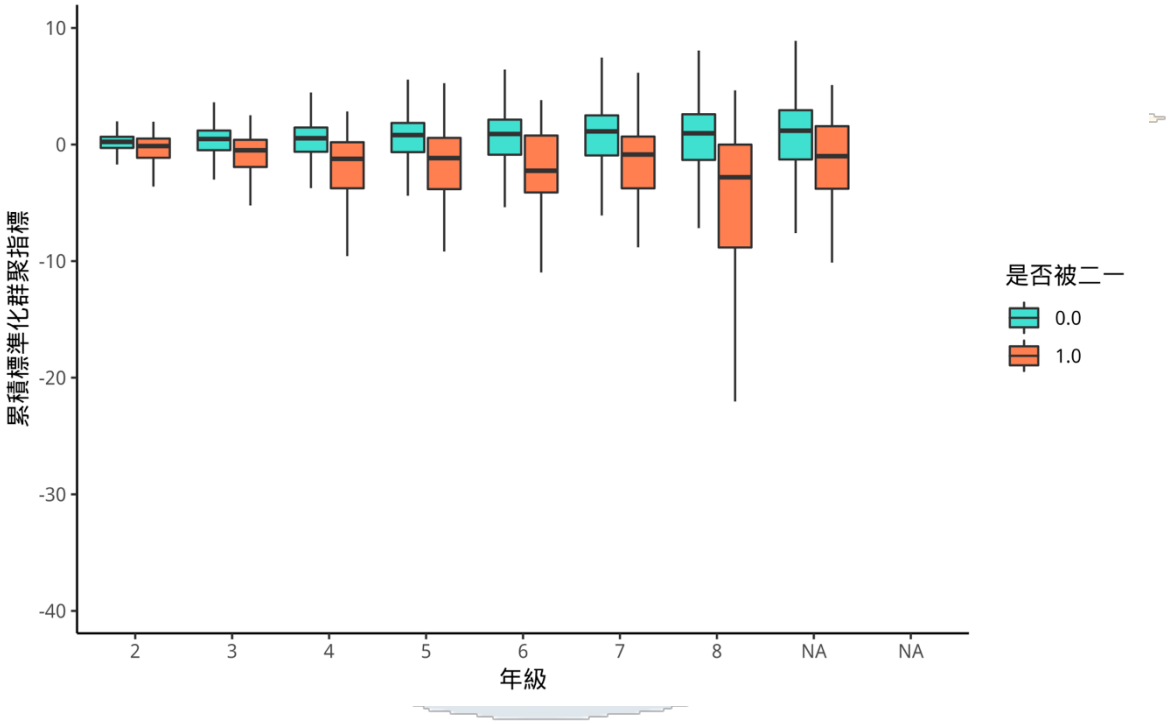
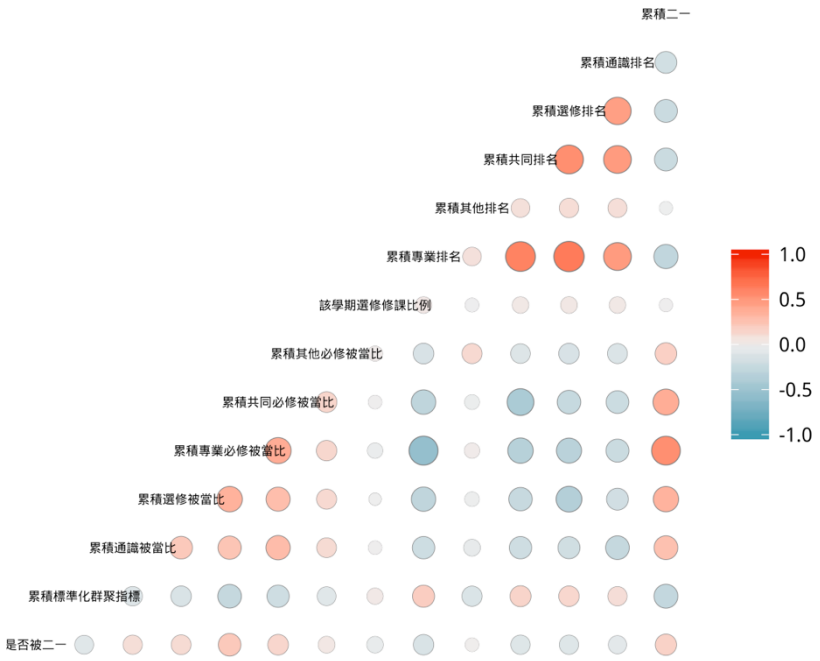


圖 2-11 顯示隨著年級上升兩類型學生的差異有越來越大的趨勢，主因為被二一的學生到後期必須去補足以前被當的科目而會導致與同班同學的課脫節，進而影響此特徵。

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。

圖 2-12：修課比例與學生二一關係圖



由圖 2-12 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

對於學生 i，給定預測二一特徵x_i，要如何產生預測結果$\widehat{待}_i$，不同預測分類工具可從「模型設定」、「模型估計訓練方式」、及「預測使用方式」來區分說明。

第二節 模型

本文將使用四種機器學習方法對學生被二一的狀況進行預測，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，以下依依介紹：

第一項 羅吉斯迴歸 Logistic regression

模型設定：

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

模型估計訓練方式：

估計方式為使用最大概似估計法，其求極值之目標函數為下

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

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

預測使用方式：

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq \hat{c} \\ 0 & \text{otherwise.} \end{cases}$$

$\hat{c}$為門檻值，若預測出的機率高於此門檻值便顯示為被二一，若低於則為沒被二一。

第二項 決策樹

模型設定：一連串的樹狀決策結構，從根結點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。

example：

“過去是否有二一”

- 是：預測會二一
- 不：預測不會二一

模型估計訓練方式：

每一筆資料依決策樹分類後，會落在其中一個葉結點，因此每個葉結果會搜集到一群資料。完美的分類必需是每個葉結點資料群都是同類的，即都被二一或都不被二一，因此，對於一棵決策樹可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。

預測使用方式：

圖 3-1：決策樹第一次分裂圖

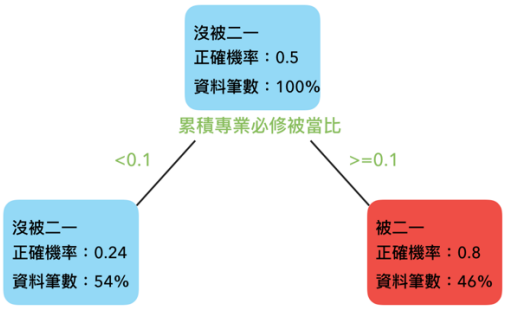


圖 3-1 中，右側葉節點的正確機率為 0.8，代表此葉節點的資料中，有八成的資料為被二一資料，模型經過不純度計算之後發現已夠低，便會停止往下分裂節點。左側之子節點裡，資料中沒被二一的比例僅有 0.24，故會繼續往下伸出節點，直到節點中的不純度夠低為止。

圖 3-2：決策樹第二次分裂圖

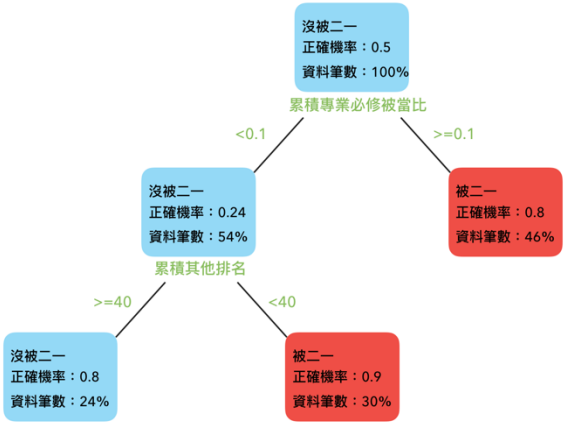


圖 3-2 中，左側子節點，繼續向下分裂，分裂出的兩個葉節，且不純度都以達到夠低，此時將停止分裂。

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多少棵樹，Hoerl A.E. and Kennard. R.W（1970）個數推薦約 64~128，假設為 K 棵樹。

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隨機抽取 K 個樣本集，並且取後放回，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為該類。

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，也可從總變數中隨機抽取 q 個變數來當作此樹的分割變數，造成樹之間的差異性。

第四項 支持向量機

模型設定：

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

模型估計訓練方式：

若訓練資料可被特徵變數超平面完美區隔成 2 類，則為線性可分，其訓練過程在於：

$$\begin{aligned} &max_{w,b} && margin(w,b) \\ &s.t. && y_i(w^T x_i + b) > 0 \text{ for } i \in S_{train} \end{aligned}$$

若訓練資料不可被特徵變數超平面完美區隔成 2 類，則為線性不可分，會利用 kernel function 於下式中以 K 表示，此時訓練過程為：

$$\begin{aligned} &max_{w,b} && margin(w,b) \\ &s.t. && y_i(w^T K(x_i + b)) > 0 \text{ for } i \in S_{train} \end{aligned}$$

預測使用方式：

$$\hat{y}_i = \begin{cases} 1 & \text{if } f(x_i|w,b) = w^T x_i + b \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

若 $f(x_i|w,b) = w^T x_i + b \geq 0$ 則預測為被二一。

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

模型設定：

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

其中

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

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:

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

模型估計訓練方式：

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

第三節 資料集

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 疊交叉驗證集來訓練超參數。 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_i ，超參數與訓練資料沒有直接關聯屬於設定變數，參數在訓練時會不斷修正而超參數並不會改變；如本文所使用支持向量機中的 C, γ ，而如何得到更好得超參數一個方法為運用驗證集來調整，由訓練集訓練不同的超參數得出模型後再丟入驗證集做驗證，藉此來調整超參數。

第四節 Synthetic Minority Over-sampling Technique

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，被二一學生數有 618 位，而訓練集中共有 7978 位學生，故本文所使用資料為不平衡資料，若是以總體分類準確率為學習目標的傳統分類演算法會將預測重點放於多數類，從而使得少數類樣本的分類效能下降，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。本文使用 Synthetic Minority Over-sampling Technique 簡稱為 SMOTE（Chawla,Bowyer,Hall and Kegelmeyer,2002）方法以解決此問題，其出發點為，減少沒被二一類的樣本，並且增加被二一類樣本數。

增加被二一類樣本的建法步驟如下：

1. 對於被二一類中每一個樣本點 y_i ，將各特徵以歐氏距離[2]為標準，分別計算它到被二一類樣本集 $S_{被二一}$ 中所有樣本的距離，然後得到每個樣本點中最近的 5 個樣本點。
2. 對於每一個被二一類樣本 y_i ，從其 5 個近鄰中隨機選擇 1 個樣本，假設選擇的近鄰為 $\hat{y}_i$ 。
3. 對於每一個隨機選出的近鄰 $\hat{y}_i$ ，分別與原樣本按照如下的公式構建新的樣本，最後將所有新合成的樣本點加入資料中。
4. $y_{i,new} = y_i + rand(0,1)(\hat{y}_i - y_i)$

   而沒被二一類資料點則是採隨機丟取，直到與被二一類新的資料筆數相同為止。

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，兩者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

[2] 歐氏距離：$Distance_{1,2} = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})}$，n 為各特徵值。

列理由，準確度會因為不同的閥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，在陽性資料與陰性資料差距很大時 AUC 擁有著比準確度還好的評價功能，例如當一筆資料中陰性類的樣本佔大多數時，面臨所有陰性樣本預測成功而陽性樣本預測失敗的情況，其準確度仍可以以達到很高，這代表了在這筆資料中準確度忽略了陽性樣本的重要性，而 AUC 可以解決此類問題，針對分類完美的模型 AUC 可以給出較高的評價，而對於僅能成功預測出單一類型的模型給予較低的評價。

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

以下將 True Positive、True Negative、False Positive、False Negative 簡稱為 TP、TN、FP、FN 。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，以下將介紹 AUC 衡量指標為何。

表格 3-1：混淆矩陣

實際 / 預測	二一	沒二一
二一	TP	FP
沒二一	FN	TN

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，ROC 曲線橫軸為誤查率，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

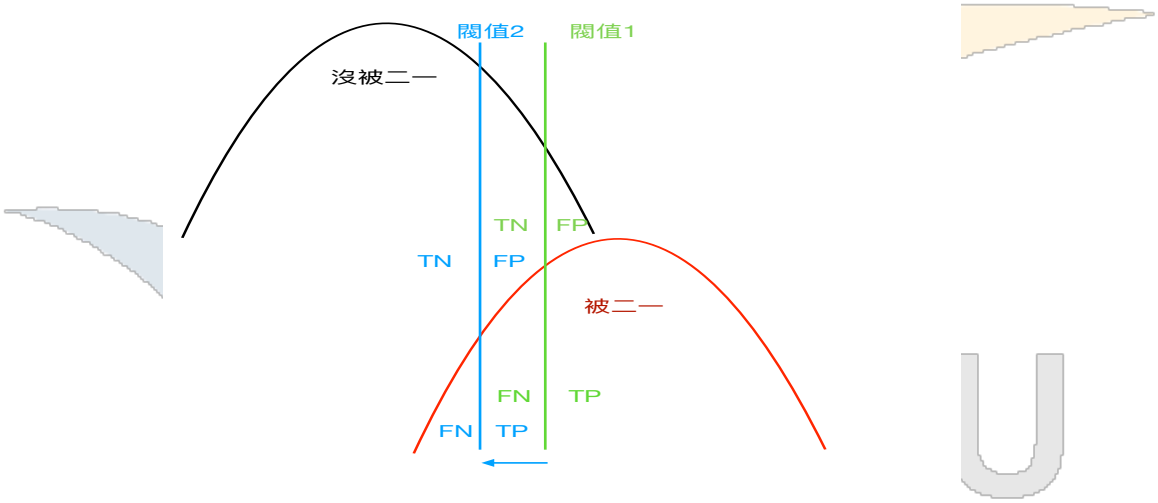
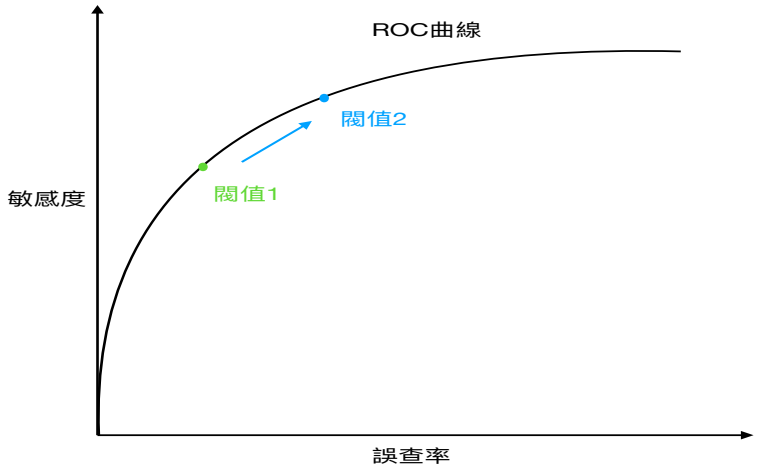


圖 3-4：ROC 關係圖



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。

第四章 研究結果

第一節 模型結果

分別比較四個模型的 AUC 與觀察混於訓練集的混淆矩陣：

圖 4-1：模型 AUC 比較圖

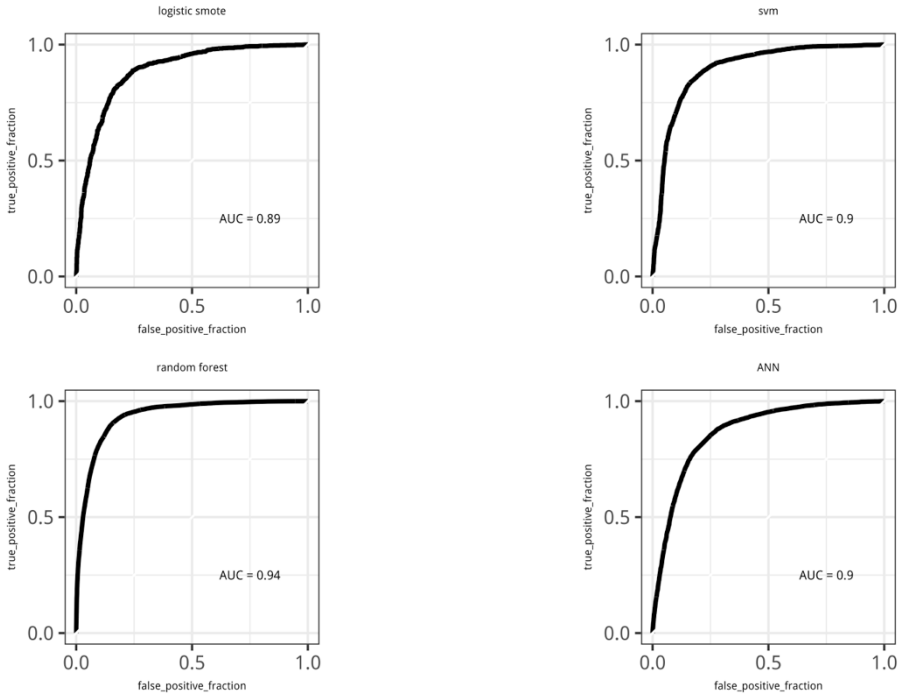


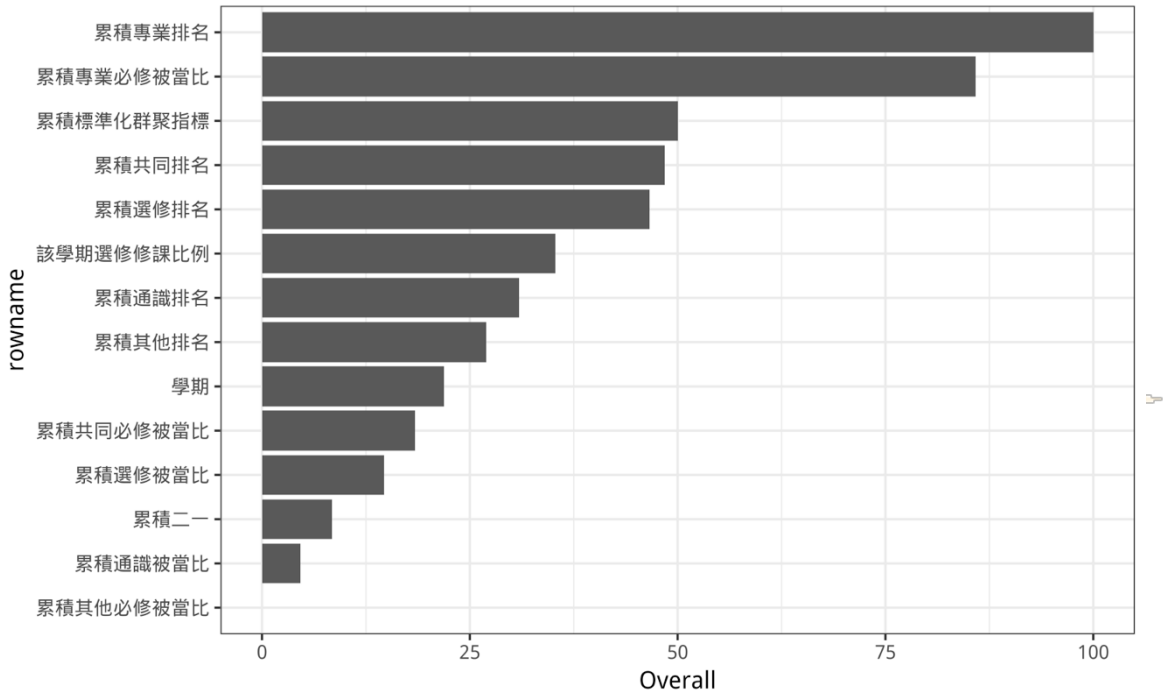
圖 4-1 中顯示，AUC 中最高的為隨機森林模型，其值高達 0.93，故將使用隨機森林於測試集資料中進行測試。

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 4-18 所示，觀察解釋能力前三的變數，累積專業必修排名、累積專業必修被當比、累積選修排名；最為重要，而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，而從圖 4-2 得知，累積專業必修排名、累積選修排名分別與累積共同必修排名具有高度的正相關，這也表明學生於大學生活上的適應狀況與專

業領域的培養之間擁有正向的關係，故若可以提高學生在專業領域的表現或許也可以使學生在大學的適應狀況提升。

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排名。本文僅用變數累積專業排名，製作羅吉斯迴歸，作為基礎模型，分別比較在測試集上的混淆矩陣。

基礎模型於測試集資料進行預測後，結果如表格 3-2。隨機森林誤查率僅有 0.16，Balanced Accuracy 即 $\frac{(\frac{TP}{TP+FN}+\frac{TN}{TN+FP})}{2}$ 為 0.82，二一學生捕捉率也高達 0.81，皆比基礎模型預測的效果來得好。

在隨機森林模型中，實際被二一的同學，模型可以成功預測出他被二一的比例高達 81%左右，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。

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	149	1788	151	1682
沒二一	37	8704	35	8810

第四節 過度擬合討論

本節探討各模型過度擬合(over fitting) 的問題，在模型訓練過程中會使用 10 折交叉驗證集的方式訓練，以防止過度擬合，且隨機森林造樹時，也會以 Bagging 的方式產生不同的個體，也可以有效的防止過度擬合的問題。

第五章 結論與建議

第一節 結論

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早好；Balanced Accuracy 為 0.815，召回率為 0.787，觀察變數間的重要性時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，而校方該如何提出一個完善的措施來提升學生專業知識的表現，改善學生的二一情形是學生與校方的重要課題。

第二節 建議

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

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 。

Vincent Tinto(1982). Limits of theory and practice in student attrition. Journal of Higher Education 53, p. 687-700.

Alaa El-Halees(2008). Mining Students Data to Analyze Learning Behavior: A Case Study.

Jiawei Han and Micheline Kamber(2006). Data Mining: Concepts and Techniques, 2nd edition.

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.

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.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy(1996). Advances in knowledge discovery and data mining.

R. S. J. D. Baker and K. Yacef(2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-16.

A. AL-Malaise, A. Malibari and M. Alkhozae(2014). Students' Performance Prediction System Using Multi Agent Data Mining Technique. International Journal of Data Mining & Knowledge Management Process (IJDKP) , vol. 4.

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.

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.

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

Pyke, S. W., and Sheridan, P. M., (1993). Logistic regression analysis of graduate student retention.Canadian Journal of Higher Education, 23, 44 – 64.

Fletcher, J., and Stren, R., (1992). Discussion of the factors influencing time to completion in graduate programs: Student views. In C. Filteau (ed.), Graduate

Graduation Rates and Time to Completion: Colloquium Proceedings (pp. 17-48). Toronto: Council of Ontario Universities.

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。