您好，我是鍾佑偉，在此報告中，我要使用的資料集是在UCI上的定期存款認購預測

那資料集就是來自UCI的銀行行銷資料

我針對此一專案所做的程式碼，網址如下，我也製成Qrcode檔，各位面試官可以透過手機掃瞄以網頁方式連結到我放在nbviewer上的jupyter notebook

而在此專案下，我主要使用了Python進行資料分析，視覺化與機器學習演算法建模

在此資料集的問題設定，是來自一家葡萄牙銀行的直接行銷活動數據，該直接行銷活動是指透過電話進行行銷銀行的定期存款，而有沒有因此而認購定期存款則成為我們的預測目標。

該資料筆數總共有41188筆

特徵量則有21筆

並不具備缺失值

而在以下這些資料是關於各個特徵所擁有的獨特值數量 以及前五個獨特值的例子

我會對各feature進行簡單介紹

Age代表目標的年紀

Job代表目標的工作型態

Martial為婚姻狀態

Education代表教育背景

Default代表是否有違約過

Housing代表是否有房子

Loan代表是否有貸款

Contact代表是用什麼裝置聯絡

Month代表是在什麼月份聯絡

Day\_of\_week代表是在星期幾

Duration代表是在最後一次接觸使用時間

Campaign代表是在這個campaign與客戶聯繫了多少次

Pdays與上次聯繫差了多久 999代表從未聯繫過

Previous這次campaign前與客戶聯繫過幾次

Poutcome之前的campaign是否有成功

環境變數

Emp.var.rate: 雇用變換率 (以季為單位)

Cons.price.idx: 消費者物價指數 (以月為單位)

Cons.conf.idx: 消費者信心指數 (以月為單位)

Euribor3m: 三個月到期的歐元銀行同業拆放利率 (以日為單位)

Nr.employed: 在職員工人數

在此，因為這個資料集是關於電話行銷，故電話行銷人員可以透過電腦螢幕來得知關於即將撥打客戶的資訊，當然，我們必須考量到人事成本，去做電話撥打的管控，避免不必要的浪費，所以會顯示建議撥打或不建議撥打，而剛剛資料集中的特徵，也都從該頁面中被提取出來。

此為在資料集下的是否認購定期存款，大約有88.7%的人會選擇不認購，11.3%的人會選擇認購，由此一分布狀況可得知，此為一個相較其他資料集起來比較不均衡的資料集，在最後的情況中，我也會針對該不均衡的問題進行調整。

此為受試者的工作型態視覺化圖表，從左邊的計數表中可得知，在行政助理，藍領以及技術人員佔了大多數，而0是未認購，1是有認購，則透過右邊的圖則可得知不同的工作種類有著多少比例的認購率，而中間橫線為剛剛圖表中整體的認購平均值，在行政人員，退休人員與失業人員和學生在此一直接行銷活動下的認購率都是高於平均值，又以學生與退休族群的認購率最高，以學生來說，通常並沒有薪水收入或者薪資收入並未非常可觀，也未對市場有所暸解，所以會對風險較低的定期存款認購有著濃厚的興趣，退休人員則因退休金穩定或無退休金，因未來無工作導致也會對理財觀念相對保守，而藍領階級在認購定期存款的比例比平均低，我認為可能是因為收入就不穩定，所以比較偏好持有現金。

每個人所受的教育程度都可能有些許的不同，文盲比例大概佔了總測驗人數的200分之一，而受大學教育的人為最多，但僅比平均高一點點的認購率，可能原因是因為在受到完整大學教育的情況下，會使人對理財有著更全面的認識，會懂的去衡量風險，並且去分散風險，調整自己的資產配置，對認購定期存款會有更謹慎和更全面的考慮。

不同的聯絡方式表示受試者在不同接聽電話的狀況下，我們可以區分出在不同的裝置下表現出的認購率，可以分析手機與家用電話通話內容或長短或接聽機率的差別，也可以根據不同的聯絡方式進行廣告的推送，

那這個活動的時長可以根據資料集的提供，是在3月到12月，在撥打次數上，明顯可以發現五月的撥打量衝到了最高，可以根據資料推估，在三月剛開始時會讓人們因為剛推出的活動或覺得新奇等等變加入認購，五月的低點可能為繳稅月，沒有錢去周轉給定期存款，9月和10月的高峰和12月的高峰可以從環境變數上面去暸解，會在之後的slides說明

在一週間不同的時間點播打電話的效果也有差，每天的撥打量都大致相同，該資料集並沒有提供在當天的幾點播打該電話和負責電話人員的id，只能依據該日期去推估可能情境，一五在客戶心情上面會比較心浮氣躁，禮拜一會有blue Monday，而禮拜五則會迎接假日的心態，但是僅僅只有些微的不同

觀看前一次campaign的使用者表現，當然還是以沒有聯絡過的客戶為主要的客群，而在上次失敗的客戶中，在這次的campaign中表現比平均還好，可以推測這次的campaign的優惠或廣告活動比上次還要強勁，或者是跟整體的景氣有關，而之前成功過的，在這次的campaign中依舊對新campaign非常捧場

受試者的年紀可以根據所做的職業稍微推斷他們的職業，但根據工作型態和年紀的資料比較下，有最高49歲的學生，也有年紀20歲的創業家，所以可以推斷也許年紀也是會影響是否認購定期存款的依據，不同的年紀段可能會對未來的規劃有著不同的想法，可以先從左圖得知，大致呈現一個右尾分配，而在中間地帶，也是不認購的比例較高，再看右圖，到30歲前大致都高於總平均，而30到60歲則低於平均，60以後則比平均高

故此，我使用在特徵工程中著名的分箱方法，嘗試將剛剛推測出來的結論去分成各個group，所以在個別的group中，能和明顯地表達不同年齡層的差異

Pdays代表與上次聯繫差了多久，999代表是之前從未聯繫過，而最高的間隔天數是27天，而在右圖當中，可以看出並不會發現一特定趨勢，間隔天數的增加並不會因此增加或減少認購的比率

在此狀況下，我也透過分箱方法將間隔天數分成兩種模式，分別為有打過，跟沒間隔天數的兩種型態，這種分箱將該變數改變成是否有撥打下一次

從該密度圖可以得知，景氣的好壞對人們認購定期存款有著重大的影響，彼此之間也以各種關係聯繫著

以利率為例，可以看出隨著時間的走勢發現，與認購率呈現高度負相關的，給了我們很好的警示以及助力應該要如何在艱難的狀況下同樣讓顧客買單，在好的狀況下讓顧客買更多的單

此為最終透過one-hot encoding後使用來建立機器學習模型的資料集，此為範例資料

在預測與實際產生錯誤的情況下，會同時有偽陽性和偽陰性的問題，偽陽性是指在預測正確的情況下，實際的值為錯誤，而偽陰性是指在預測錯誤的情況下，實際的值為正確，而在此一專案下，偽陽性是代表我在預期他會訂購定期存款下對其進行電話行銷，而最後從電話中得知該客戶並不會認購我們的產品，而偽陰性則是在打電話前就認為他是不會訂購的，就不會對其進行電話行銷，但其實如果對其進行電話行銷，他是會認購定期存款產品的，所以根據上述情況，我們必須降低偽陰性的結果，因為少一個人認購產品的損失會大大超過客服人員撥打一通電話的成本，故我們要降低閥值，讓越來越多的會被判定為正確，雖然有可能提高偽陽性發生的機會，但是卻可以降低錯失潛在客戶的風險，具體的閥值計算必須透過衡量這兩種結果的成本來得知。

羅吉斯回歸是使用線性的方式並透過sigmoid的方法將其轉成機率的性質，並透過剛剛上述所說的必須讓false negative最小化的方法，降低打電話的門檻，也有將其進行特徵縮放中的常態化，並使用l2的方法當作懲罰項，避免羅吉斯回歸有過擬和的現象產生。

故前一張投影片所說，因為是線性，所以每個變數皆會有其權重，而對最終預測皆會有正向或負向的影響，可以從第一個變數看起，是教育中的文盲，表示如果其為文盲，將大大增加其認購的機率，再來則是我們進行加工過的previous day，意指為之前有聯絡過的，也會增加認購機率，同樣的，之前的成果結果，和在三月的表現都大大增加認購的可能性，而對其有負面影響的則是雇用數量，已經在五月進行行銷活動

而下一個使用的機器學習演算法是隨機森林演算法，在這邊，在此我改用h2o來進行機器學習演算法的訓練，因為scikit-learn的算法中並不支持categorical參數的訓練，一定得使用one-hot-encoding來處理特徵，而h2o不用，他可使用categorical的方式來處理，所以接下來的隨機森林決策樹與梯度上升決策樹都是使用h2o進行訓練和預測。而我選定的閥值為0.07，但也回有超過9000筆資料被模型判定為不認購，而我則盡可能去提高recall的分數，讓成本最小化

在決策樹中的變數重要性意思為該變數對模型能分辨二元的變數貢獻了多少的比例，在此途中，我們可以看中，環境變數是影響人們做決策的一大變數，而再來就是人們的工作型態和教育程度，活動聯繫的次數，與在星期幾聯絡

最後的模型則是梯度上升決策樹，模型準確定為0.74，因為必須降低閥值的關係，所以會使得準確率降低，但是在recall的score卻相對較高，recall的意思是在實際正確的強況下，預測正確的機率

而在gradient boosting的重要性下，可以得出環境變數依舊在前三名，而月份，工作型態，星期幾和教育程度都對模型有顯著的貢獻