資料探勘期末報告

第 14 組

組員

309709034 林蔚恩
A101228 徐佳靖
0713137 洪翊鈞
0853721 楊浩智
309709025 劉珊妮

中華民國110年01月15日

一、題目簡介與動機：Bank Marketing

本次專題題目為Bank Marketing，由於金融機構的家數及分行數量眾多，且銀行的同質性非常高。各銀行所提供的產品幾乎是相同的金融商品，因此要從劇烈競爭中勝出並不容易，若銀行要在存款性的金融機構中競爭，更須隨時調整行銷的策略與技巧。

銀行主要收入為銀行的存放利差 (Spread) 與服務費收入 (Fee)，若資產品質管理不好，或行銷不當，就會發生重大虧損。銀行的創新與行銷方式提升，才能有永續存在與成長的機會。此數據集說明葡萄牙銀行機構的直接行銷活動也就是電話行銷之相關內容，主要的分類目標是預測客戶是否會購買定期存款（y)，行銷方式主要在於客服電話撥打。通常，需要與同一客戶有多次聯繫，以便尋問客戶是否會購買銀行定期存款。因此期望能夠透過機器學習與深度學習之方式預測客戶是否會購買定期存款，並期望能夠針對不同的客群運用不同的電話行銷方式來進行產品推廣，以提升客戶購買定存產品的數量。

二、資料介紹

銀行客戶資訊特徵

特徵名稱	資料型態	資料說明
Age	integer	客戶年齡
Job	object	職務類別（admin, blue-collar, entrepreneur 等，共12種）
Marital	object	婚姻狀況（divorced, married, single, unknown ，共4種）
Education	object	教育程度（basic.4y, basic.6y, basic.9y, high.school等共8種）
Default	object	是否信用違約？（有、無、未知）
Housing	object	是否有房屋貸款？（有、無、未知）
Loan	object	是否有個人貸款？（有、無、未知）

客戶聯繫相關特徵

特徵名稱	資料型態	資料說明
Contact	object	與客戶聯繫方式（cellular, telephone）
Month	object	聯繫月份（jan, feb, mar, ..., nov, dec）
Day_of_week	object	聯繫星期（mon ,tue, wed, thu, fri）
Duration	integer	聯繫時長，聯繫結束前為未知，且為 0 時預測特徵為No，因此需把這個特徵從輸入特徵中移除

經濟情況特徵

特徵名稱	資料型態	資料說明
Emp.var.rate	float	就業變動率 - 季度指標
Cons.price.idx	object	消費者價格指數 - 月度指標
Cons.conf.idx	object	消費者信心指數 - 月度指標
Euribor3m	object	Euribor 3 個月利率 - 每日指標
Nr.employed	object	員工人數 - 季度指標

其他特徵

特徵名稱	資料型態	資料說明
Campaign	integer	在活動期間聯繫此客戶次數
Pdays	integer	從上一個活動中最後一次聯繫客戶後經過的天數
Previous	integer	在活動之前和客戶聯繫的次數
Poutcome	object	上次行銷活動的成果（失敗、成功、未知）

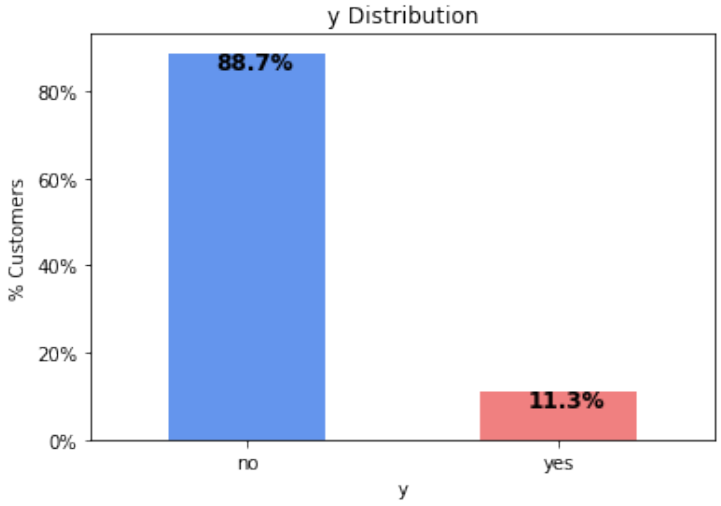
預測特徵

特徵名稱	資料型態	資料說明
y	object	客戶是否會購買定期存款（會：yes 、不會：no）

在這個資料集中，總共包含了21個特徵，我們選取了一個當作預測特徵，其他20個作為備選的輸入特徵。另外，資料集總共包含了41188筆資料，並且沒有遺漏值。

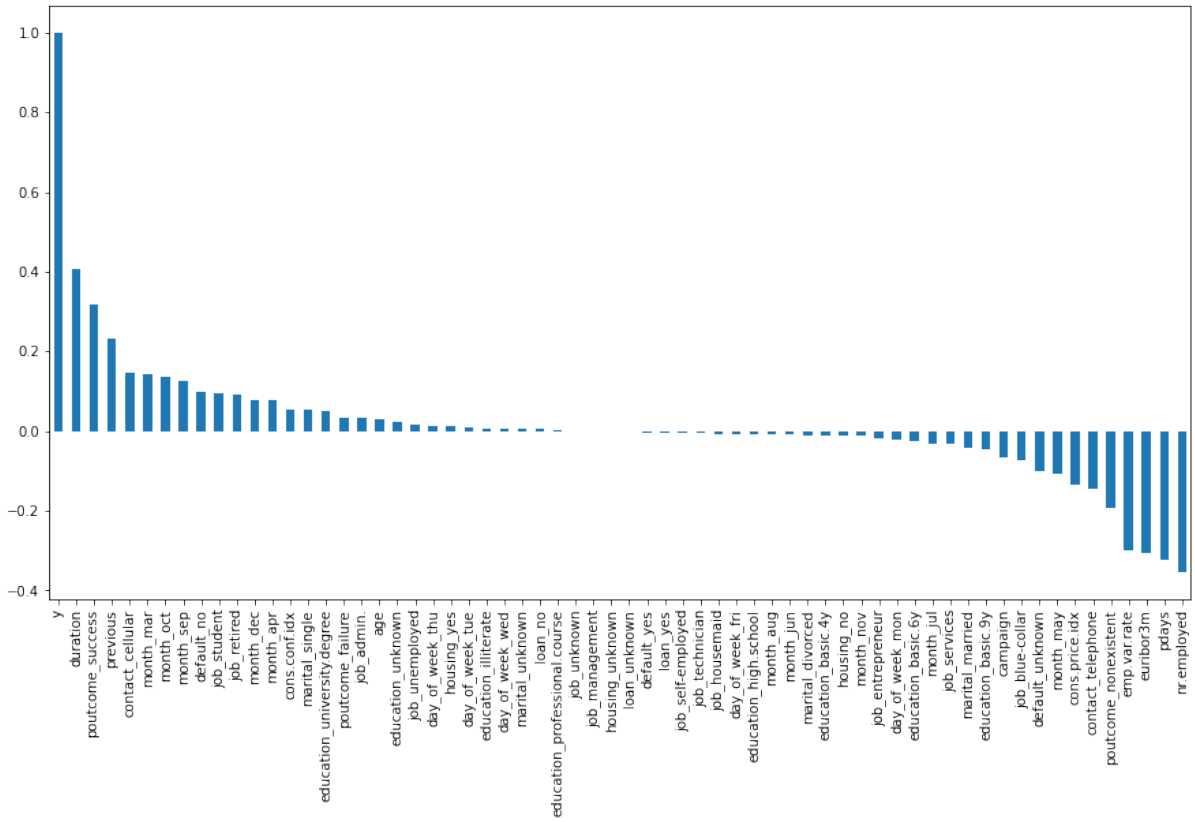
三、探索性資料分析 EDA

1.是否購買定期存款



目標變數「是否購買定期存款」，以虛擬變數呈現，有購買的客戶佔樣本資料11.3%，沒有購買的客戶則佔比88.7%，如上圖所示。

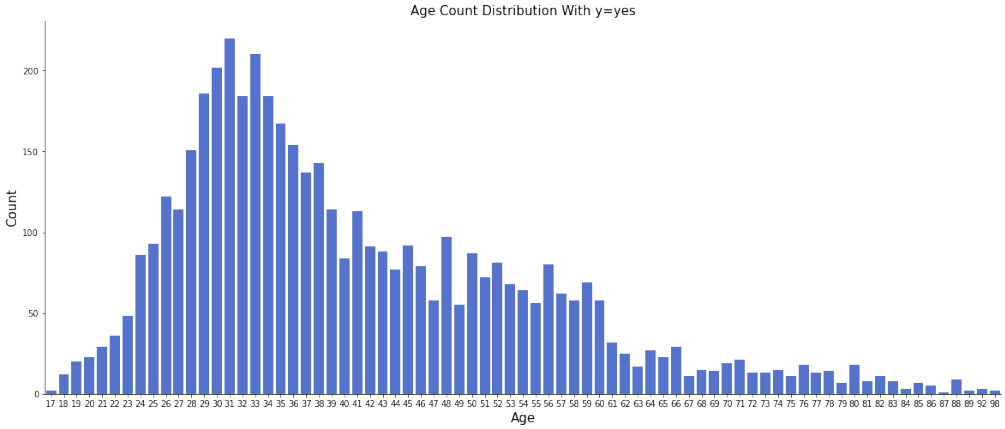
2.相關性分析



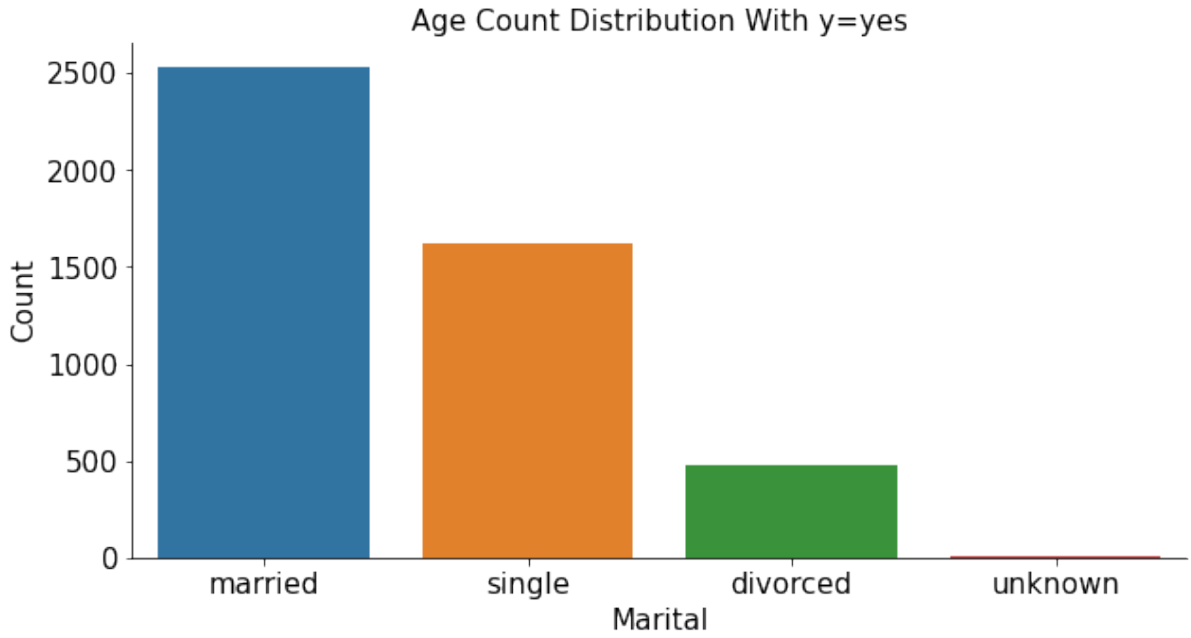
上圖中呈現的是y這個特徵為yes時，將類別變數轉為虛擬變數後，與所有變數之間的相關係數，透過這個相關係數圖我們可以發現y與duration、poutcome-success、previous、contact-cellular和聯繫的月份之間出現正相關。而與員工人數、Pdays、Euribor 3 個月利率和就業變動率之間呈現較為明顯的負相關。因此我們後面在進行探索性資料分析時可以多觀察這幾個特徵，甚至把他們加入到預測模型中。

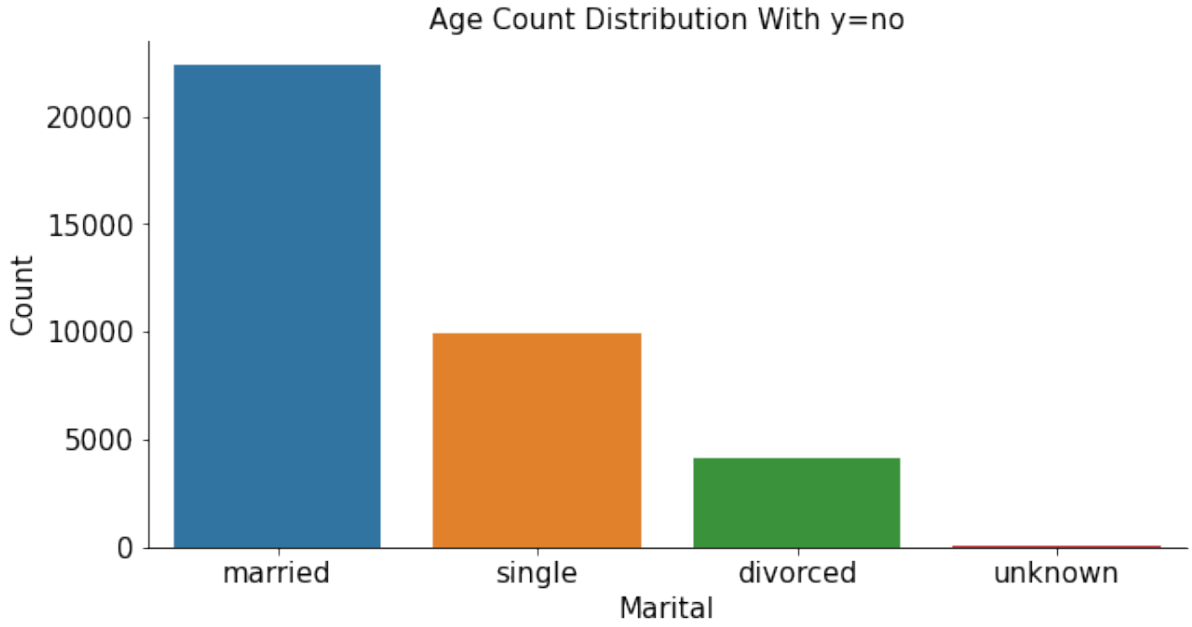
3.年紀

由下方的年齡分佈圖可以發現，有購買銀行定期存款的客戶(y=yes)年齡分佈呈現右篇分佈且較為集中在25到40歲，因此若要預測客戶是否會購買，年齡的分佈可能為一個重要特徵。



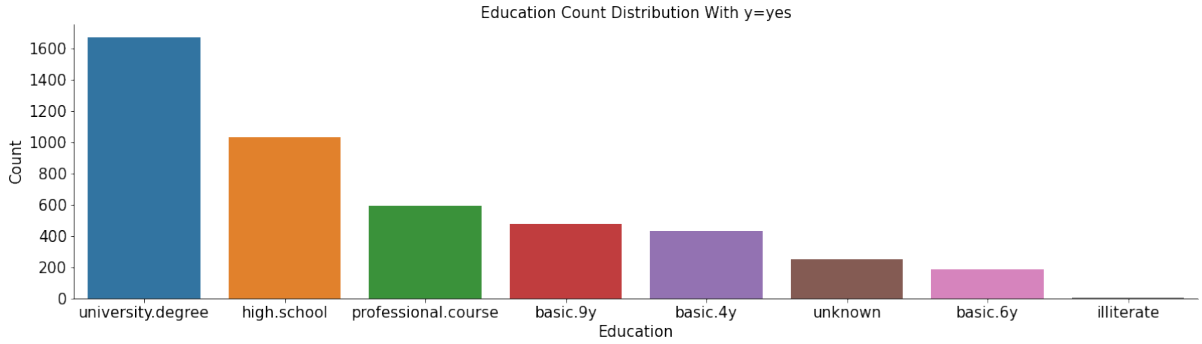
4.婚姻

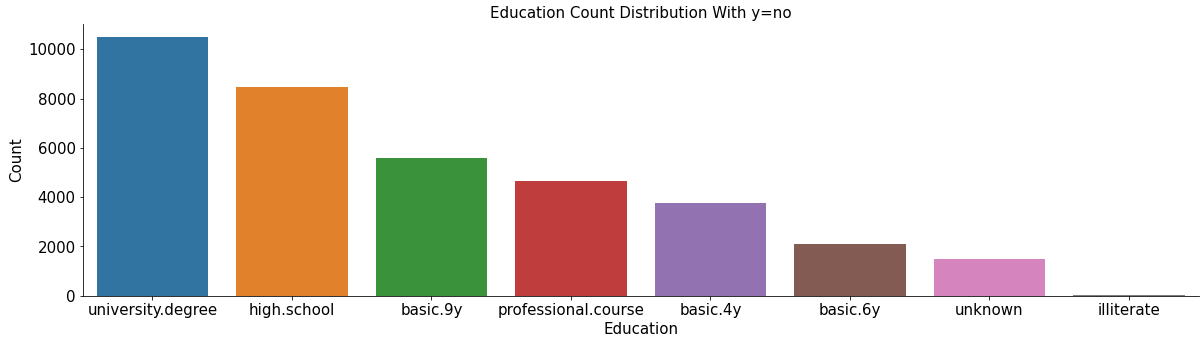




根據上述有購買與沒有購買定期存款的客戶婚姻狀況分佈皆為已婚者為居多，而離婚者之分佈較少，因此可得知婚姻狀況對於有無購買定期存款的類別來說差異性並不大，因此婚姻狀況在分類上並沒有太大的意義。

5.教育

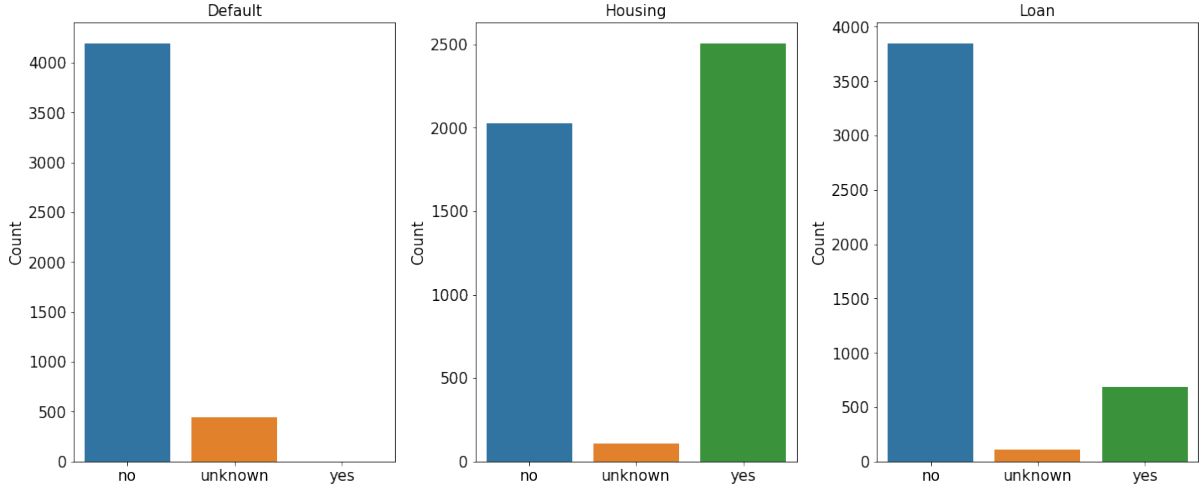




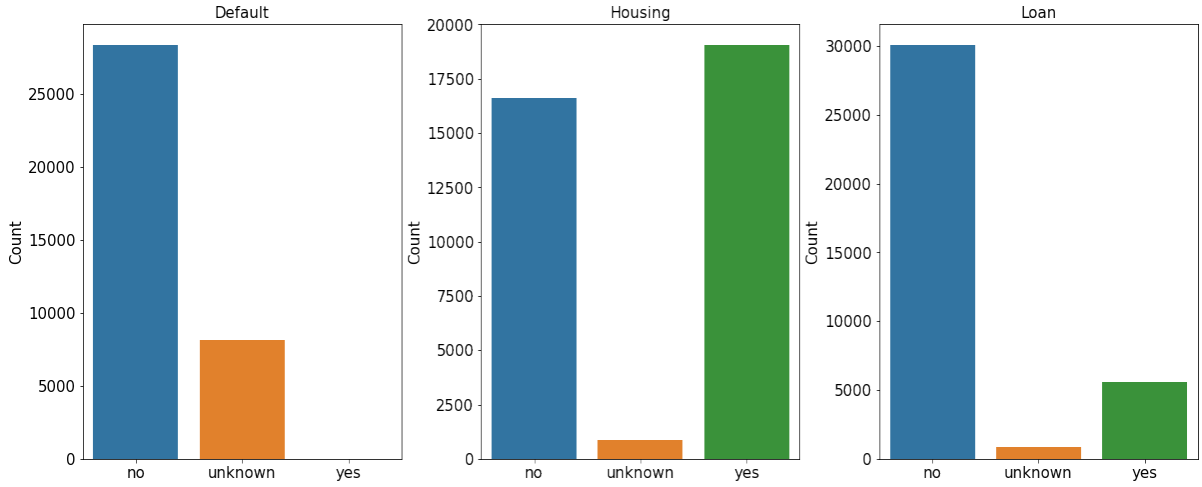
有購買定期存款之客戶與沒有購買定期存款之客戶最高學歷皆為大學，而次高學歷皆為高中，但在有購買定期存款之客戶第三高的學歷為碩士，沒有購買定期存款的客戶第三高的學歷為國中，因此可得知有購買定期存款之客戶在教育程度上可能比沒購買定存的客戶較為高，教育程度在分類上可能為一個重要特徵。

6.違約、貸款和房地產

y=yes



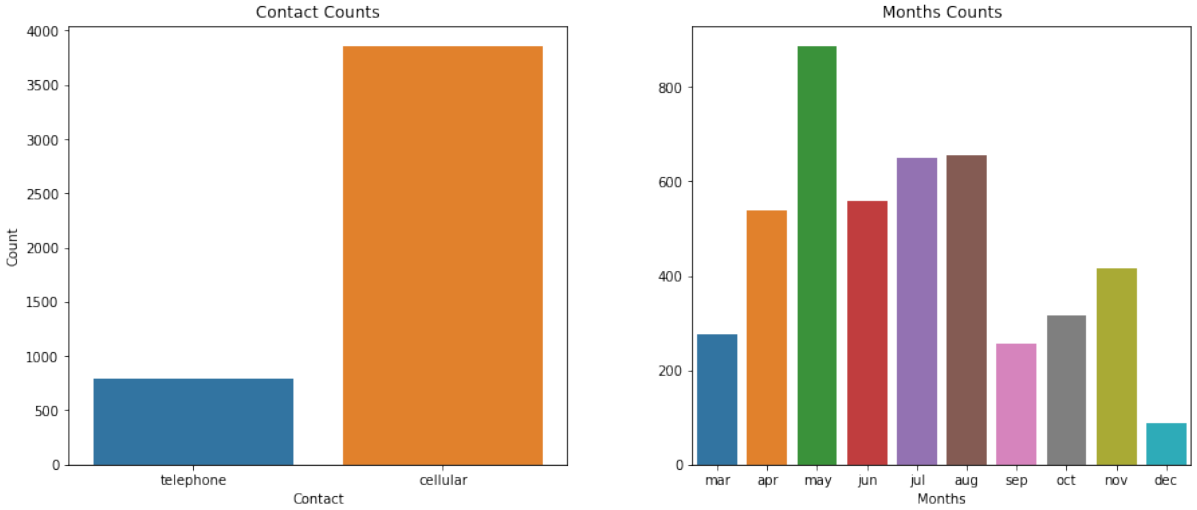
y=no



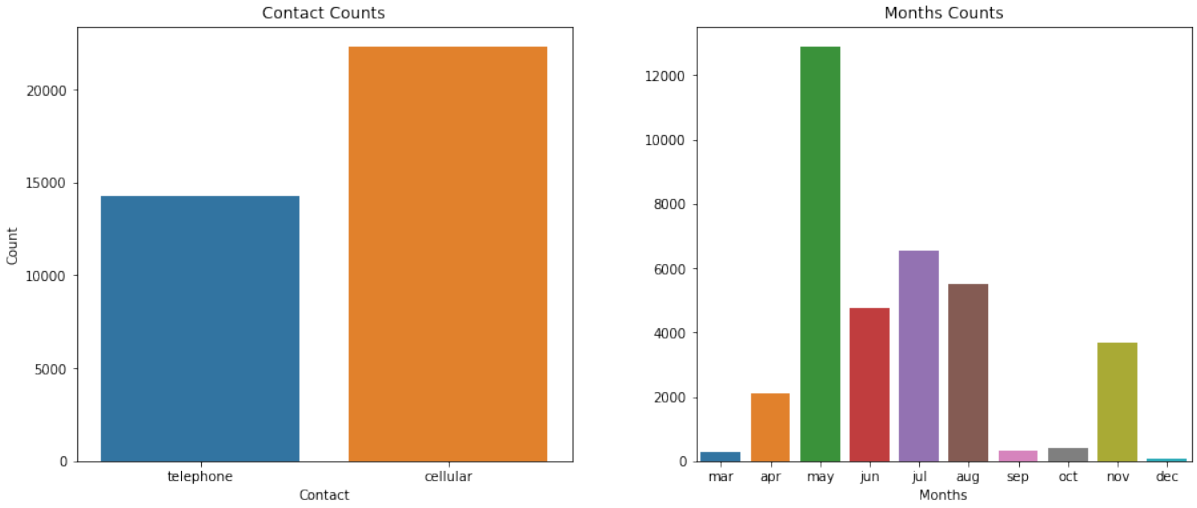
在違約的分佈上有購買定存的客戶數量（4000)明顯少於沒購買定存的客戶有購買定存的客戶(25000)在資產管理方便可能較為警慎且信用較為良好，因此可以透過違約這個特徵來進行分類。而在房地產與貸款的分佈在這兩類客群上較為雷同。

7.聯繫方式、聯繫月份

y=yes



y=no



聯繫方式若是以家用電話的方式來聯繫客戶會導致購買定存的客戶比例較少，用行動電話聯繫客戶較可能讓客戶購買定存，因次聯繫方式為重要之特徵。

而在聯繫客戶的月份上，在三月、九月、十月與十二月份期間聯繫客戶，購買定存的客戶的比例較沒有購買的比例為高。

四、資料前處理

後續建立的模型會分為淺層模型以及深度學習兩類，因此在資料前處理也會有不同步驟。

1. 淺層模型

從下圖中可以看到，本次期末專題選取的資料集並無缺失值，因此不需要進行補值。由於任務型態是二元分類問題，因此將購買定期存款(y=yes)設為1；不購買定期存款(y=no)設為0。進行資料前處理之前，先將原始資料以4:1的比例分為訓練以及測試資料集，以避免資訊洩漏。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

(1) 特徵工程

下表是數值資料的敘述統計，從第一、第三分位數和最大最小值可以看到，campaign、pday和previous存在離群值，推測可能含有大量相同的數值，因此須檢視特徵值的分布狀況；從第三分位數和最大值來看，age、duration存在離群值。

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000
mean	40.038574	257.643490	2.570319	962.837967	0.171533	0.079253	93.574997	-40.509627	3.619153	5166.991266
std	10.448502	257.101849	2.787879	186.013340	0.491837	1.572500	0.579550	4.639493	1.734857	72.191324
min	17.000000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.000000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.000000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.000000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.000000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

A. Campaign: 活動期間與客戶聯繫的次數

下圖是聯繫次數依照頻率排序的部分結果，1次、2次、3次占了81.5%左右，與客戶聯繫3次以上的頻率並不高，因此將此特徵進行離散化處理，分成1次、2次、3次和3次以上。

1	0.426980
2	0.258270
3	0.129681
4	0.063460
5	0.038998
6	0.024279
7	0.015417
8	0.010015

B. Pday:從上一個活動中最後一次聯繫客戶後經過的天數

999的比例佔了96.4%左右，代表本次活動聯繫的客戶絕大多數在上次活動沒聯繫過，因此將此特徵進行離散化處理，0代表上次活動過後並未聯繫過客戶；1 代表上次活動過後聯繫過客戶。

999	0.963581
3	0.010592
6	0.009833
4	0.002792

C. Previous:在活動之前和客戶聯繫的次數

可以看到0和1的比例佔了97%左右，因此進行離散化處理，分為活動開始前與客戶聯繫過0次、1次和1次以上。

0	0.864279
1	0.110076
2	0.018422

D. 數值形態特徵

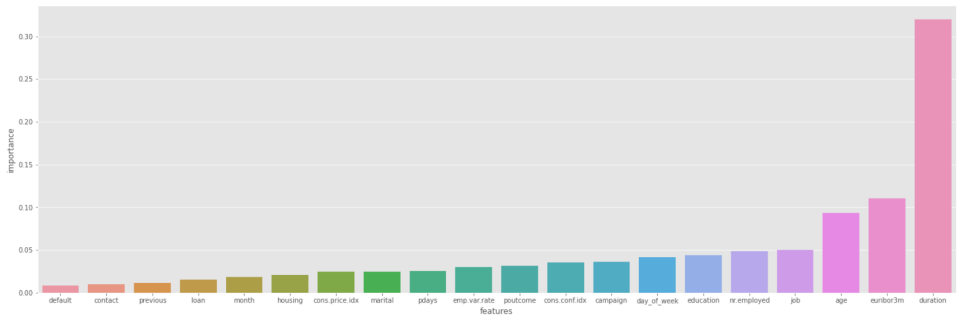
由於age和duration存在離群值，但不想刪掉離群值而損失資訊，因此使用RobustScaler進行標準化，再不刪除樣本的情況下減少離群值的影響，確保建模時的穩健性。

E. 類別形態特徵

針對淺層模型進行處理，不用像類神經網路一樣將類別特徵進行OneHot Encoding，因此以OrdinalEncoding轉換類別資料，也可以避免稀疏性問題。

(2)特徵選取

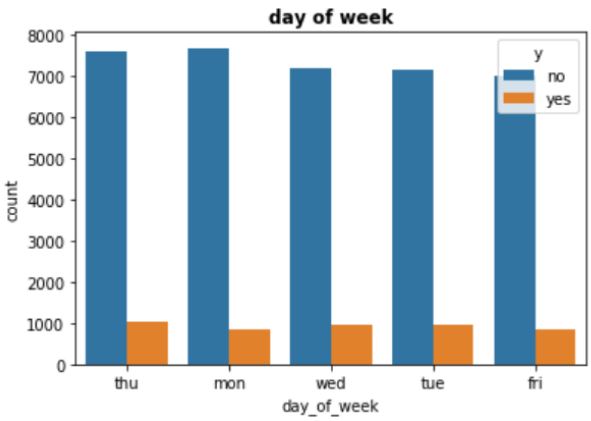
希望選取80%的特徵(16個)，但是使用Backward Selection搭配隨機森林的方式，實驗了幾次發現每次選取的特徵都不一樣，因此先訓練隨機森林模型，刪除Gini importance對低的四個特徵，分別是: default、contact、previous和loan。

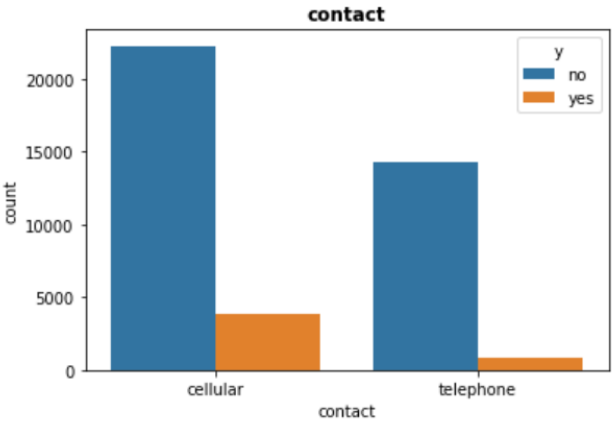


2. 深度學習

本份資料無缺失值，並不需要進行補值，此環節將聚焦於移除多餘欄位

1.移除和y關聯不顯著contact、day_of_week





如圖所示，可發現contact、day_of_week兩欄位對於y類別的影響不大為了加速模型運算以及提升準確率，將contact、day_of_week兩欄位移除

轉換類別資料

利用pandas內建的get dummies 函數，將類別資料進行轉換，以0、1進行表示

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	...	month_dec	month_jul	month_jun	month
0	56	261	1	999	0	1.1	93.994	-36.4	4.857	5191.0	...	0	0	0	
1	57	149	1	999	0	1.1	93.994	-36.4	4.857	5191.0	...	0	0	0	
2	37	226	1	999	0	1.1	93.994	-36.4	4.857	5191.0	...	0	0	0	
3	40	151	1	999	0	1.1	93.994	-36.4	4.857	5191.0	...	0	0	0	
4	56	307	1	999	0	1.1	93.994	-36.4	4.857	5191.0	...	0	0	0	

五、建模

1.淺層模型

本次選用了以下4個模型:

A. Logistic Regression
B. Support Vector Machine
C. Random Forest
D. XGBoost

其中Logistic Regression作為Baseline model，與其他效能更強的分類器進行比較，因此Logistic Regression不進行超參數調整。而其他三個模型會使用Bayesian Optimization進行超參數調整，由於從EDA的結果可以看到，標籤y(是否會購買定期存款)存有類別不平衡問題，因此除了Accuracy外，也選用了Precision、Recall和F1 score做為模型衡量指標。

在超參數調整的部分，將訓練資料切割10%作為驗證資料，也就是以資料集72%(0.8*0.9) 的資料作為調參的訓練資料，8%(0.8*0.1)來驗證，並以最大化Accuracy作為超參數優化方向，三個模型各迭代50次，將找到的最佳超參數訓練最後的模型，並以測試資料比較其效能。

1.Support Vector Machine

調整kernel(核函數)、C(正則項)以及gamma(與支持向量數成反比)，下圖是最優超參數以及調參耗時:

```
SupportVector Machine
========================================
Hyperparameters process of costs: 146.99 s
Hyperparameters of best trial are: {'kernel': 'rbf', 'C': 166.84908720937028, 'gamma': 70.2159739690894}
```

2.Random Forest

調整n_estimators(樹的棵數)、max_depth(樹的深度)、max_features(每顆樹的特徵數)、min_samples_leaf(葉節點所含最小樣本數)和min_samples_split(分裂節點最小樣本數)，下圖是最優超參數以及調參耗時:

```
RandomForest
========================================
Hyperparameters process of costs: 818.59 s
Hyperparameters of best trial are: {'n_estimators': 100, 'max_depth': 17, 'max_features': None, 'min_samples_leaf': 30, 'min_samples_split': 43}
```

3.XGBoost

主要調整樹的深度和數目、正則化係數、學習率、節點分裂門檻，下圖是最優超參數以及調參耗時:

```
XGBoost
========================================
Hyperparameters process of costs: 1394.54 s
Hyperparameters of best trial are: {'max_depth': 15, 'subsample': 0.7666163551396739, 'n_estimators': 700, 'eta': 1.930644196751853e-08, 'alpha': 0.0465181148
798173, 'lambda': 0.020276459323665376, 'gamma': 0.019354624073880607, 'min_child_weight': 10, 'grow_policy': 'lossguide', 'colsample_bytree': 0.9666351028793
168}
```

4.調參效能比較

從下表可以看到，由於XGB生成樹時有順序性，但是RF可以平行處理，因此XGB耗時比起RF還要多；耗時最少的是SVM，但是之前的實驗過程中，若沒有限制最大迭代次數，會因為找不到最佳分割超平面使電腦當機，因此本次SVM的調參結果是限制迭代次數為1000次求得的超參數。

從RF的max_features和XGB的colsample_bytree可以看到，建模過程中，兩種算法都傾向選取所有特徵來建構子樹。

	SVM	RF	XGB
耗時(s)	146.99	818.59	1394.54

2.深度學習

建立一個有5層的神經網路，共有5552個參數，模型摘要如下

```
Model: "sequential_22"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_83 (Dense)             (None, 52)                2756
_________________________________________________________________
dense_84 (Dense)             (None, 35)                1855
_________________________________________________________________
dense_85 (Dense)             (None, 20)                720
_________________________________________________________________
dense_86 (Dense)             (None, 10)                210
_________________________________________________________________
dense_87 (Dense)             (None, 1)                 11
=================================================================
Total params: 5,552
Trainable params: 5,552
Non-trainable params: 0
```

六、模型比較

1.淺層模型

下表是根據上一節找到的最優超參數所建立的模型，應用於測試資料的各項評量指標，可以看到其實Baseline的Logistic Regression表現不錯。因為資料存在類別不平衡的問題，因此比起Accuracy，其他三項指標更為重要。

(1)Aaacuracy
所有模型的準確率皆大於88.7%，也就是資料中標籤的原始分布，其中以XGBoost的表現最好；SVM表現最差。

(2)Precision
Precision代表在模型預測會購買定期存款的客戶中，有多少比例是真正有購買的。銀行會希望可以透過模型，找出最有可能購買的客戶，並針對這些客戶投入資源，吸引客戶在銀行定期存款，以達精準行銷的目的。其中以XGBoost的表現最好，Logistic Regression的表現次之，SVM的表現最差。

(3)Recall
Recall代表在有購買定期存款的客戶中，有多少比例是模型預測正確的。其中以隨機森林的表現最好，SVM的表現最差。

(4)F1 score
F1 score是Precision和Recall的調和平均數，也是評價模型比較全面的綜合指標，尤其是在類別不平衡的資料及中更為重要。其中以隨機森林的表現最好，XGBoost的表現次之。

綜合上述分析，SVM的表現在各個模型之中是最差的，基本上找不出預測會購買且真的有購買的客戶(TP)，或許是因為限制了最大迭代次數，使模型無法找到最佳分割超平面，因此在模型訓練以及調參的過程中，或許需要學習和參考更多trick來提升模型效能。

隨機森林和XGBoost是Ensemble learning中代表Bagging和Boosting的經典算法，確實在效能上比起Baseline Model有所提升。雖然從F1 score來看，隨機森林的表現是最好的，但是實務上F1 score沒有直觀的解釋性，而且本次目標是希望透過機器學習模型協助銀行找出最有可能購買定期存款的客戶，因此Precision才是我們應該關心的重點，其中以XGBoost的表現最好，而且在其他三個指標中的表現也都很不錯，因此我們認為XGBoost是最適合作為最終預測的模型。

從本次結果來看，Logistic Regression的表現不錯，尤其是Precision的表現更是僅次於XGBoost，從運算成本以及效能來看，Logistic Regression的CP值非常高。金融業由於涉及大量金流，因此會比其他產業受到更嚴格的主管機關規範，以模型效能和預測能力的角度來看，XGBoost無疑是最佳選擇，但是也常常被詬病為黑盒子，算法過程不易被解釋，因此簡易的Logistic Regression仍是在兼顧運算成本、模型效能以及透明度的情況下，常被使用的模型。若要再提升Logistic Regression的效能，可以從資料前處理再深入鑽研，透過資料分箱(binning)或是監督式編碼，例如WOE Encoding等等，將非線性特徵轉化成Logistic Regression善於處理的線性特徵，以提升模型預測效能。

	Logistic	SVM	RandomForest	XGboost
Accuracy	0.907866	0.887109	0.914664	0.915028
Precision	0.643463	0.250000	0.639405	0.653639
Recall	0.408405	0.001078	0.556034	0.522629
F1 score	0.499670	0.002146	0.594813	0.580838

2.深度學習

混淆矩陣、f1 score、recall、precision、accuracy、花費時間如下圖

```
322/322 [==============================] - 0s 982us/step - loss
[[8365  774]
 [ 280  878]]

The accuracy is 0.8976400893464116
time cost 15.949599504470825 s
f1_score: 0.6249110320284698
recall: 0.7582037996545768
precision: 0.531
accuracy: 0.898
```

深度學習模型像個黑盒子，在解釋機器如何學習上較為困難，管理者需在模型校用以及解釋性上做取捨。以這份資料集建模的結果來看，深度學習並沒有特別的表現，反而是Random forest 和XGBoost有相對教高的準確度和精確度等的評估指標。

	Logistic	SVM	Random forest	XGBoost	ANN
Accuracy	0.907866	0.887109	0.914664	0.910528	0.897640
Precision	0.643464	0.250000	0.639405	0.653639	0.531000
Recall	0.408405	0.001078	0.556034	0.522629	0.758203
F1 score	0.499670	0.002146	0.594813	0.580838	0.624911

七、結論與管理意涵專題

研究結果可以使銀行根據過往資料更精準地找出什麼特性的消費者願意購買定期存款，精準的找出目標客群可以節省更多撥打無效電話的成本，也能夠提升顧客購買之意願，

綜合上述使用淺層與深層的五種模型，我們發現淺層模型在指標上有更好的表現，在Accuracy上以Random forest 和XGBoost獲得的0.91為最佳，在其他指標Precision、Recall和F1 score 的表現不相上下，但我們的主題更適合關注Precision 獲得的分數，這表示我們建立的模型預測會購買定期存款且最後實際有購買定期存款的比例，比例愈

高代表我們浪費的行銷資源相對少，也表示行銷的目的明確，可以為公司在節約成本的情況下創造效益。

本專題所使用的分析工具與方法也適用於其他產業，例如：電信業、保險業、補習業，上述產業與銀行行銷之手法雷同，皆須透過電話行銷來吸引消費者，因此本次專題所研究之模型與參數也可運用至這幾個產業使用，透過機器學習與深度學習之法更精準地找出目標客群；此外，這個模型的概念也可以應用在銀行的其他業務，例如信用卡推銷或貸款（車、房貸），以目前市場情況來看，貸款推銷多利用隨機電話推銷，導致消費者接到電話時常感覺到厭煩，因為消費者並不需要這些服務。從銀行推銷的角度，這樣的隨機推銷也耗費大量成本，可能也不見成效。所以，銀行在推銷信用卡或貸款服務前可以先蒐集消費者資訊，並依據這些資訊按照這個專題的研究過程，建立機器學習模型，減少耗費成本的同時，也提高推銷帶入的效益。

八、未來改進

從資料集的角度來看，可以納入其他重要特徵，如：是否有外匯資產、活期存款金額多少、目前是否已經購買定存和目前定存金額多少等的特徵，這些特徵可能會對消費者是否購買下一筆定存有很大的影響，若未考慮這些特徵可能無法精準的尋找到會購買定存的客戶。另外，也可以增加一項紀錄該名客戶之前接通這類推銷電話的時間，以推測消費者在何時是方便接到推銷電話，以目前市場情況來看，推銷電話並沒有特定會撥進的時段，導致消費者接通時當下是在上班而無空閒接通電話。

九、參考資料

https://www.kaggle.com/henriqueyamahata/bank-marketing
https://archive.ics.uci.edu/ml/datasets/bank+marketing