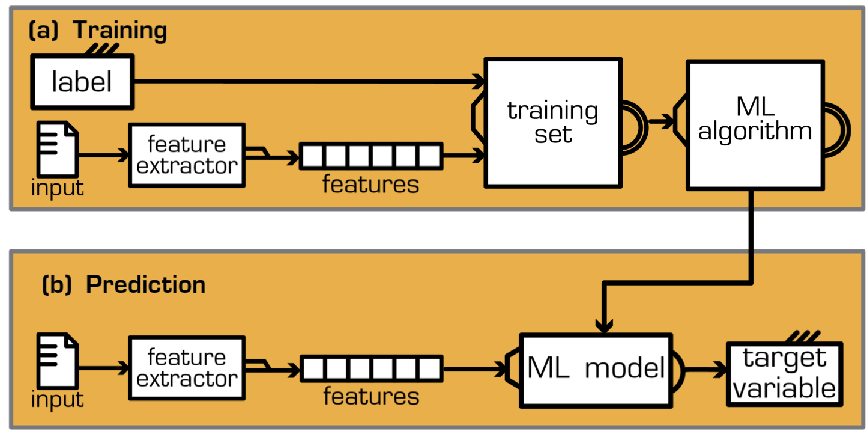
1. 研究背景與目的

在這個資訊的時代，在我們生活中許多程序都會留下數據。以本次的題目來說，銀行會提供信貸服務，但銀行只是給人貸款服務，當然也要考量營運面，若借貸人無能力還款也是種負擔。然而，如果從借貸人的數據去分析有沒有還款能力呢？畢竟數據那麼多，若靠人工運算太沒效率，因此機器學習這時候就能夠幫助許多議題進行決策，將關鍵指標放入模型進行評估以幫忙快速進行決策。

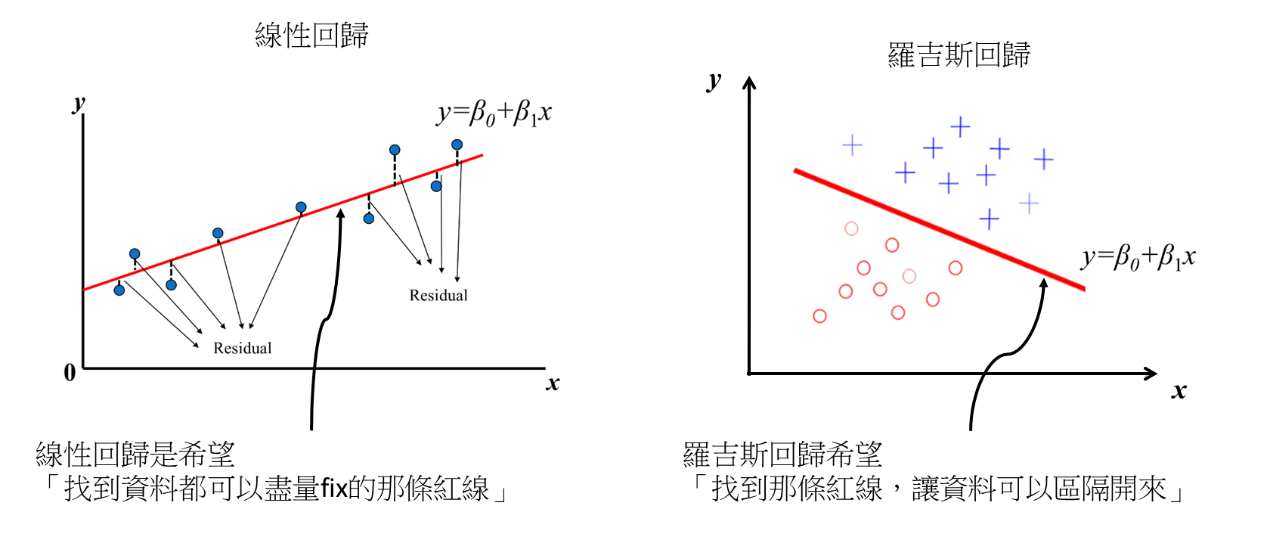
本次主要透過spark mllib classification針對先有的交易資料進行還款預測。

1. 研究方法
   1. SPARK ML PIPELINE

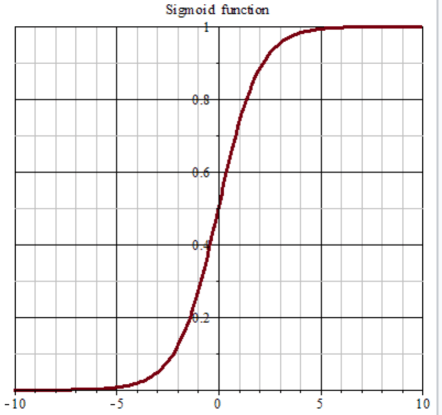


如上圖為模型訓練的架構圖，機器學習的流程通常會被不同類型的資料進行ETL，主要會根據input內容去提取特徵並分成訓練與測試，將訓練集來訓練模型並評估測試集的成效。

本次演算法使用邏輯斯迴歸。邏輯斯迴歸是處理二元分類常用的一種方式，屬於線性分類器的一種。主要針對二元問題進行分類。

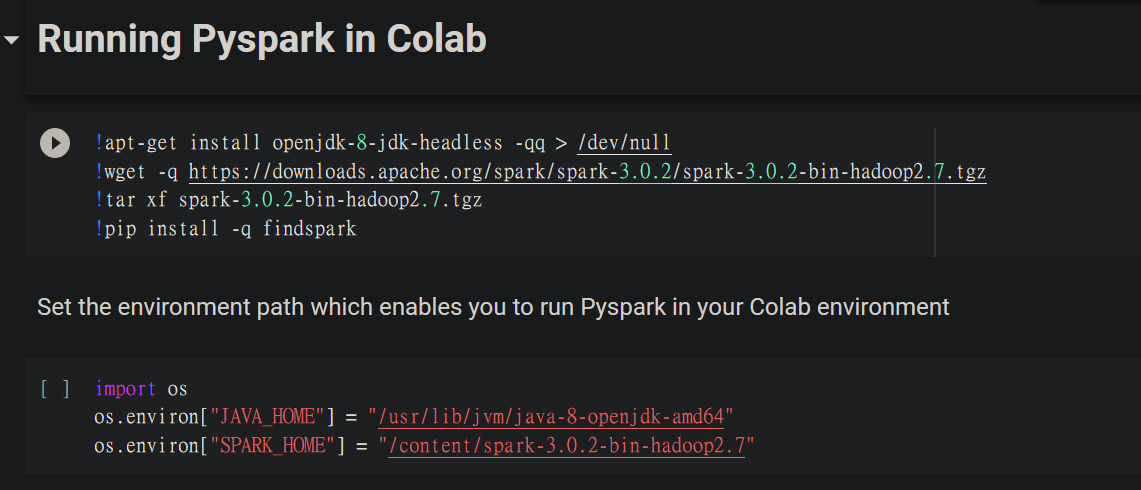


在邏輯斯迴歸模型中，還有很重要的一點就是要經過Sigmoid函數的轉換，主要的目的就是要讓最後的結果介於0與1之間，以便預測。



* 1. 環境架設

首先，建立pyspark環境。由於本次是以Colab來執行，所以需要做些事前設定。



再來，就是針對pyspark需要的套件import到專案。



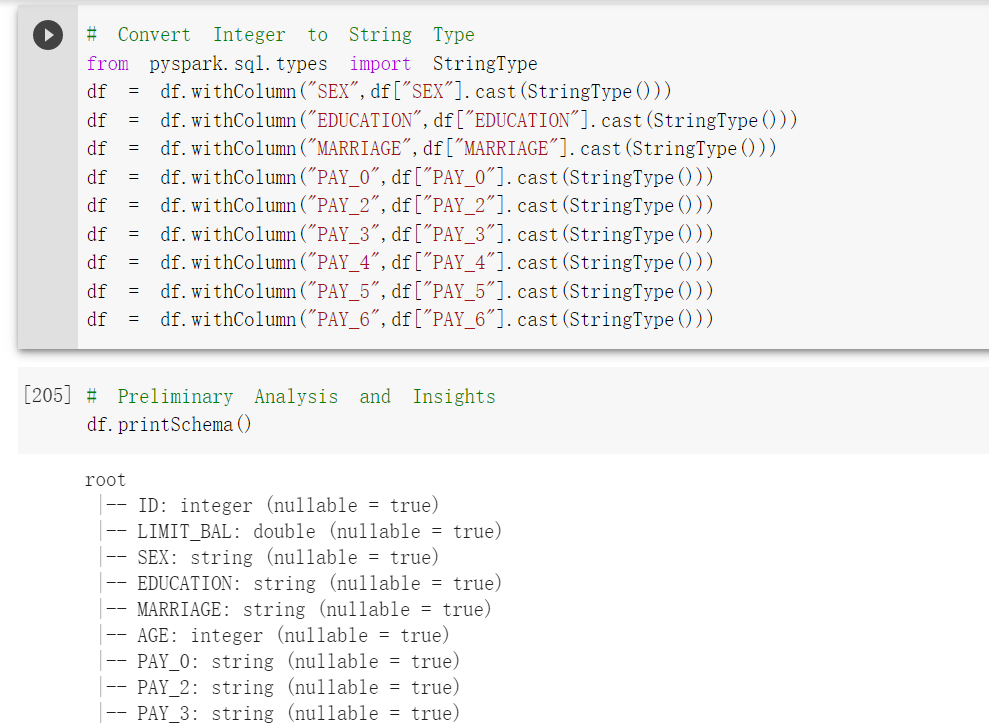
* 1. 資料探索與清理

使用資料時，我們也需要對資料作了解。如下內容為本次資料介紹：

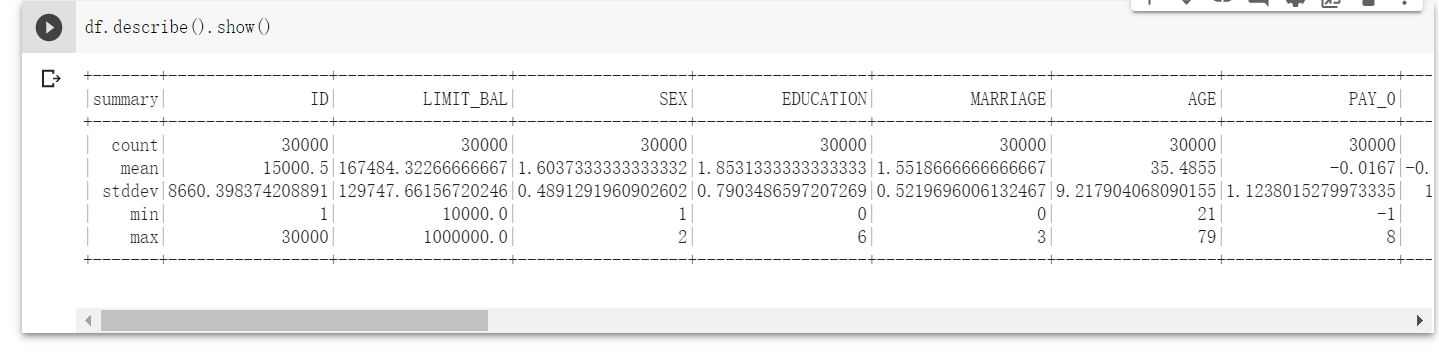
* + ID: ID of each client
  + LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
  + SEX: Gender (1=male, 2=female)
  + EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
  + MARRIAGE: Marital status (1=married, 2=single, 3=others)
  + AGE: Age in years
  + PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
  + PAY\_2: Repayment status in August, 2005 (scale same as above)
  + PAY\_3: Repayment status in July, 2005 (scale same as above)
  + PAY\_4: Repayment status in June, 2005 (scale same as above)
  + PAY\_5: Repayment status in May, 2005 (scale same as above)
  + PAY\_6: Repayment status in April, 2005 (scale same as above)
  + BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
  + BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
  + BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
  + BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
  + BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
  + BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
  + PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
  + PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
  + PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
  + PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
  + PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
  + PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
  + default.payment.next.month: Default payment (1=yes, 0=no)

從資料中，有性別、學歷、婚姻狀態、帳單和繳費資訊等信息，有些內容在資料用數值表示類別，像是學歷、婚姻、性別，因此特別了解數值的意義。

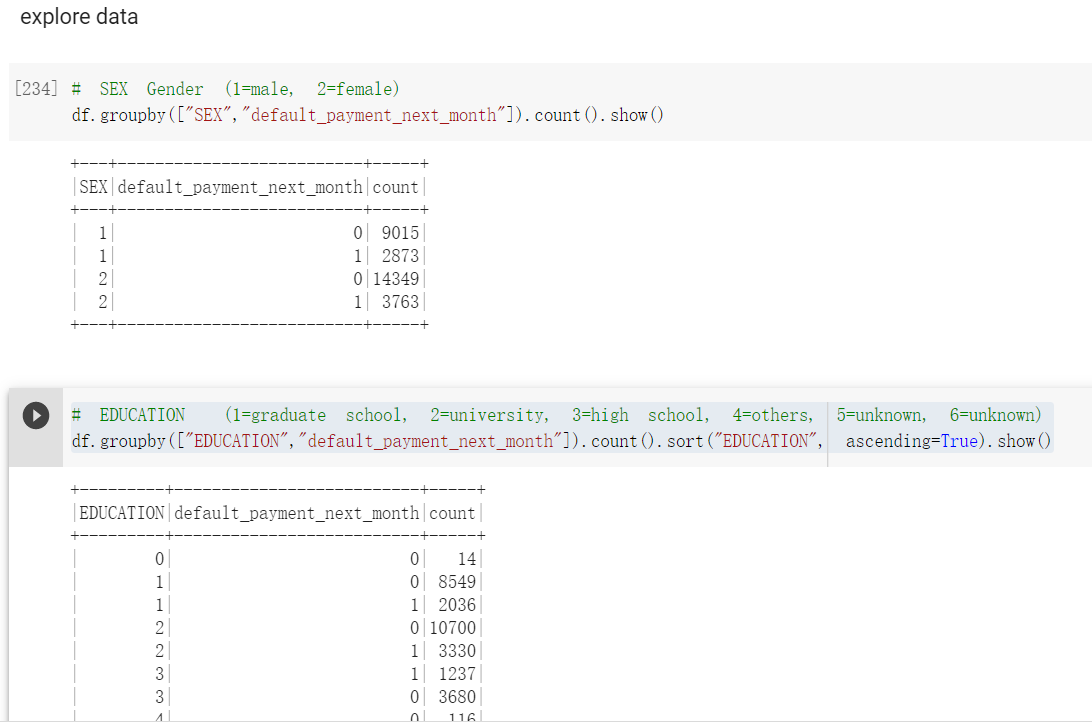
將資料讀取並觀察欄位格式，基本上都是數值型態。有些欄位並非數值為indicator，因此事先做轉換。



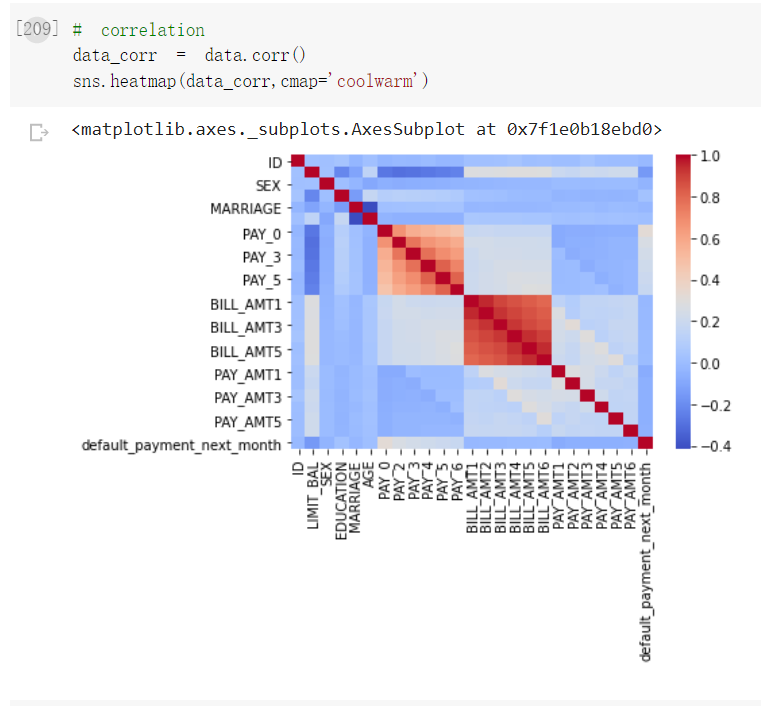
透過指令也能夠了解資料分布情形。



也能夠過指令觀察input & output之間關係。



透過視覺化圖表更能直接觀看之間的相關性，但沒有一個欄位與我們的目標特別顯著的關聯。

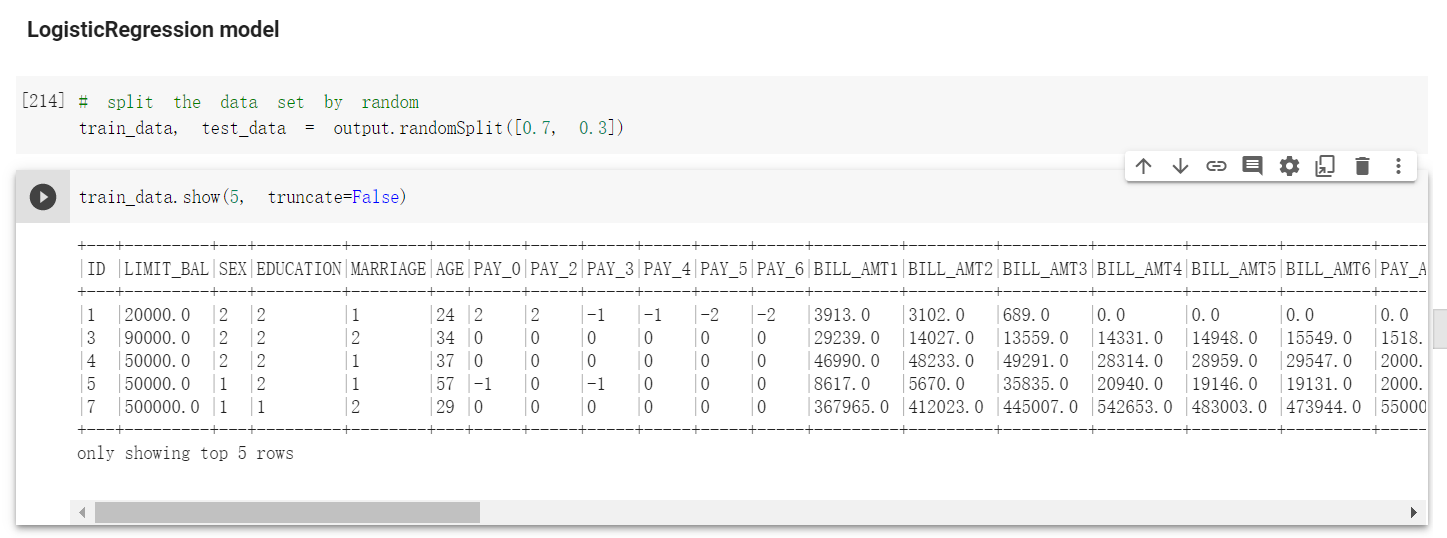


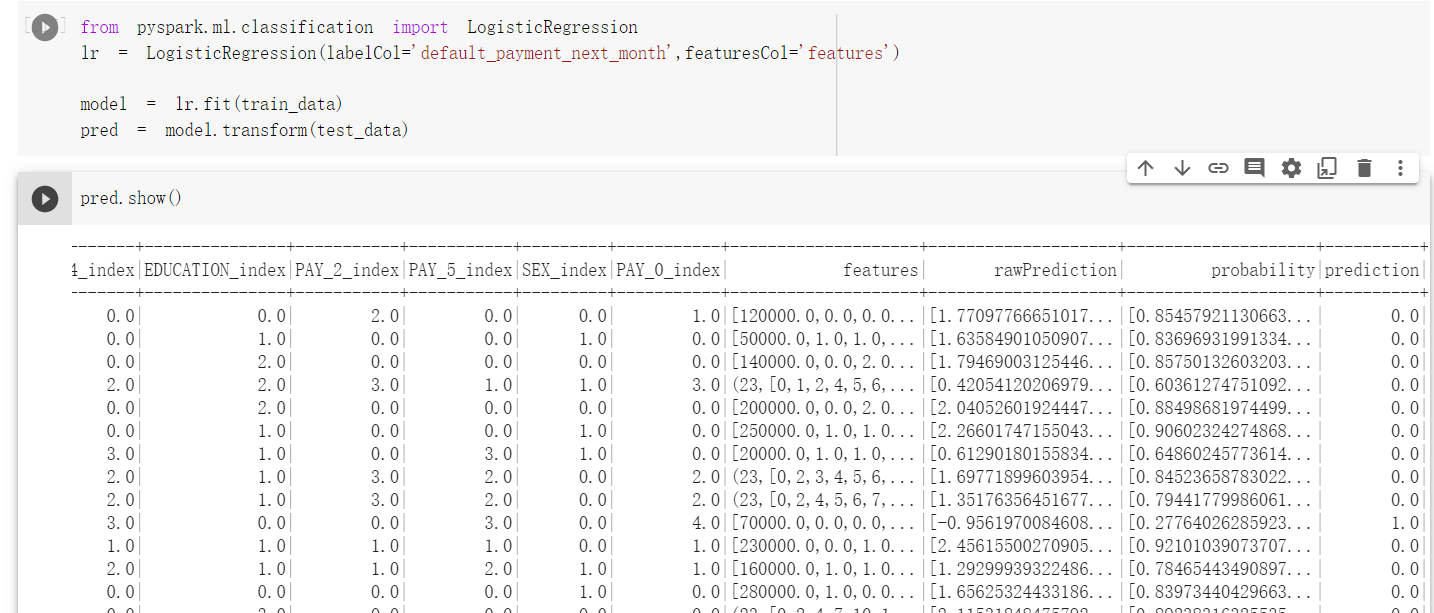
* 1. 使用pineline 進行預測

在放入模型之前，需要先收集到的特徵資料轉換。本資料有數值型和類別型，因此採用StringIndexer、VectorAssembler來解決。



特徵轉換後得到的features得放入模型進行訓練，訓練好的模型才能夠進行預測。





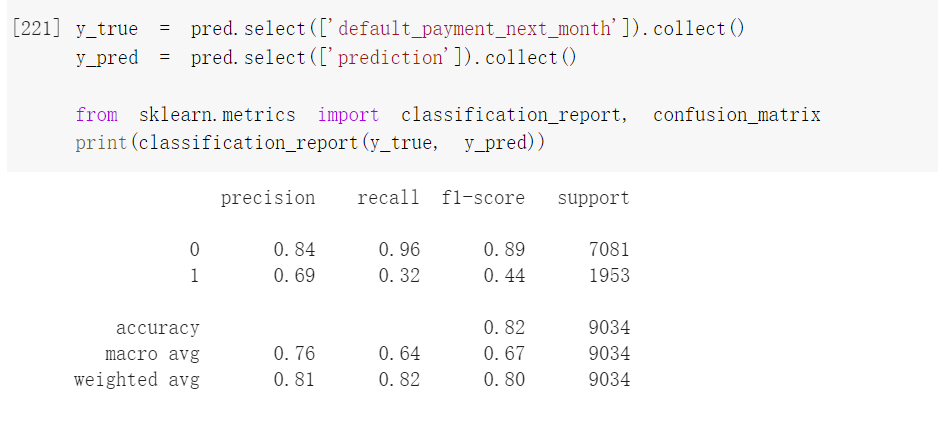
* 1. 評估模型

BinaryClassificationEvaluator為分類評估指標，獲取roc曲線面積為0.64。RegressionEvaluator為回歸模型評估指標，均方根誤差RMSE為0.42。



Precision看的是在預測正向的情形下，實際的「精準度」是多少，而Recall則是看在實際情形為正向的狀況下，預測「能召回多少」實際正向的答案。

模型結果默認不會付款的Precision為84%，會付款的Precision為69%；模型結果默認不會付款的Recall為96%，會付款的Recall為32%。模型的準確率也有82%。



Confusion Matrix是用於分類問題的一種常用的指標。能夠快速觀看預測跟實際差異。可以發現模型訓練結果針對有些內容還是較難判定是會默認付款。



小結，這次透過羅吉斯回歸模型訓練表現還行，但有些部分還是可以再加強。這是資料集可能會遇到的問題，像是目標標籤數量不平均，因此資料較少的內容可能會比較難看出來規則，未來也考慮針對特定欄位再訓練或是換演算法，說不定能夠提高會默認付款的recall。

1. 困難與心得及人員分析

在本次分析有許多心得，主要是很開心自己能夠更有成就感，學會使spark ml進行模型預測。雖然在過程中碰了不少壁，像是怎麼下語法、遇到環境安裝的問題、資料轉換或是命名規則不符之類，但慢慢查詢相關資料就慢慢迎刃而解。

1. 參考資料

1. default of credit card clients Data Set https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

2. Default Payments of Credit Card Clients in Taiwan from 2005 https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

3. Pyspark on Colab

https://www.analyticsvidhya.com/blog/2020/11/a-must-read-guide-on-how-to-work-with-pyspark-on-google-colab-for-data-scientists/ [pyspark on colab]

4. Binary Classification Example https://docs.databricks.com/\_static/notebooks/binary-classification.html

5. An introduction to seaborn

https://seaborn.pydata.org/introduction.html

6. 《巨量資料技術與應用》實務操作講義- Spark機器學習程式庫：MLlib

http://debussy.im.nuu.edu.tw/sjchen/BigData-Spark/%E5%B7%A8%E9%87%8F%E8%B3%87%E6%96%99%E6%8A%80%E8%A1%93%E8%88%87%E6%87%89%E7%94%A8%E6%93%8D%E4%BD%9C%E8%AC%9B%E7%BE%A9-MLlib.html

7. PYSPARK IN THE CONTEXT OF SALES DATA

https://www.rrighart.com/pyspark