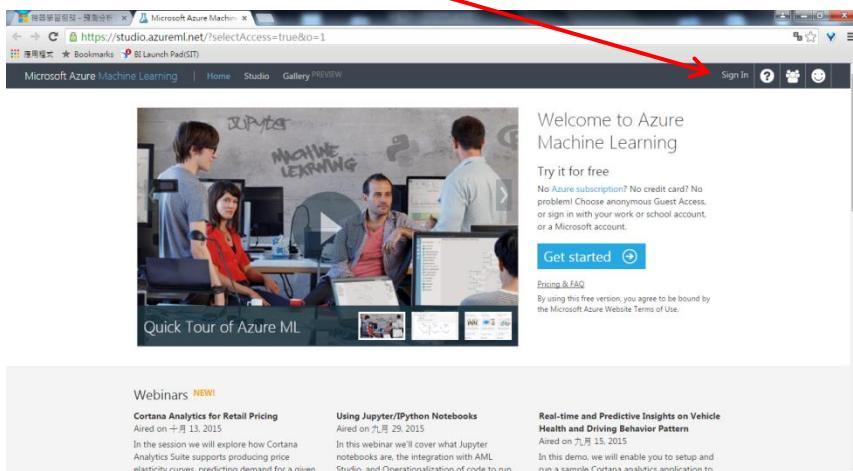


Demo for MS Azure Machine Learning:

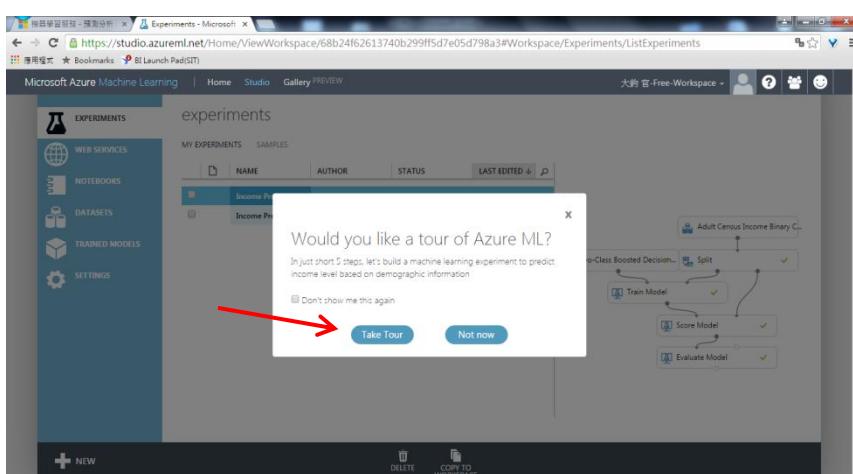
1. 登入網頁: <https://azure.microsoft.com/zh-tw/services/machine-learning/>
並選擇立即開始使用>



2. 點選右上方的 Sign In
並登入 Hotmail 帳戶

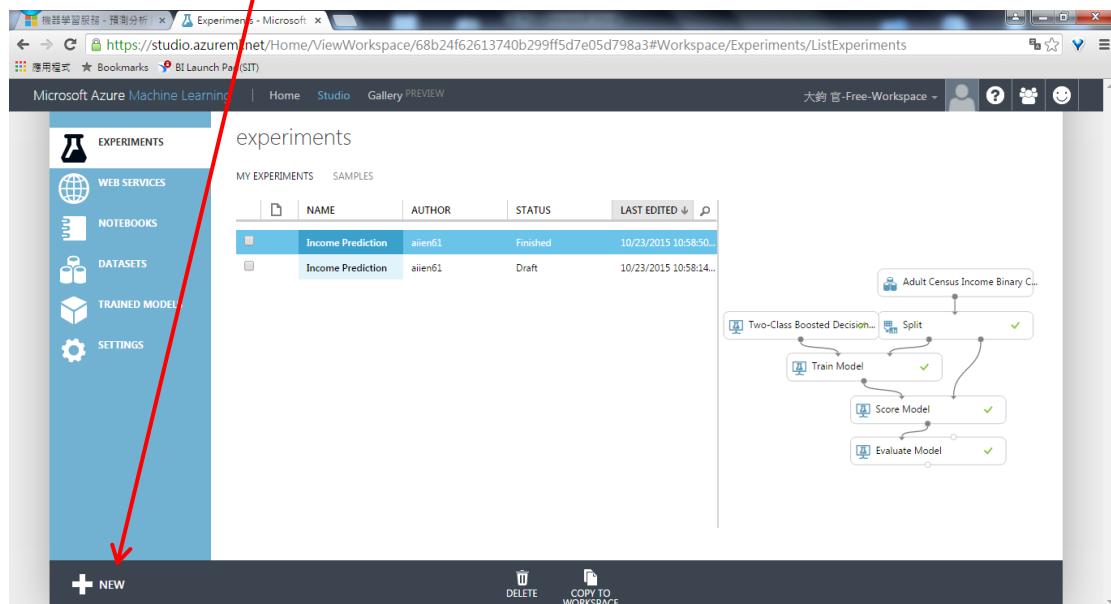


3. 可選擇 Azure 引導一個範例教學

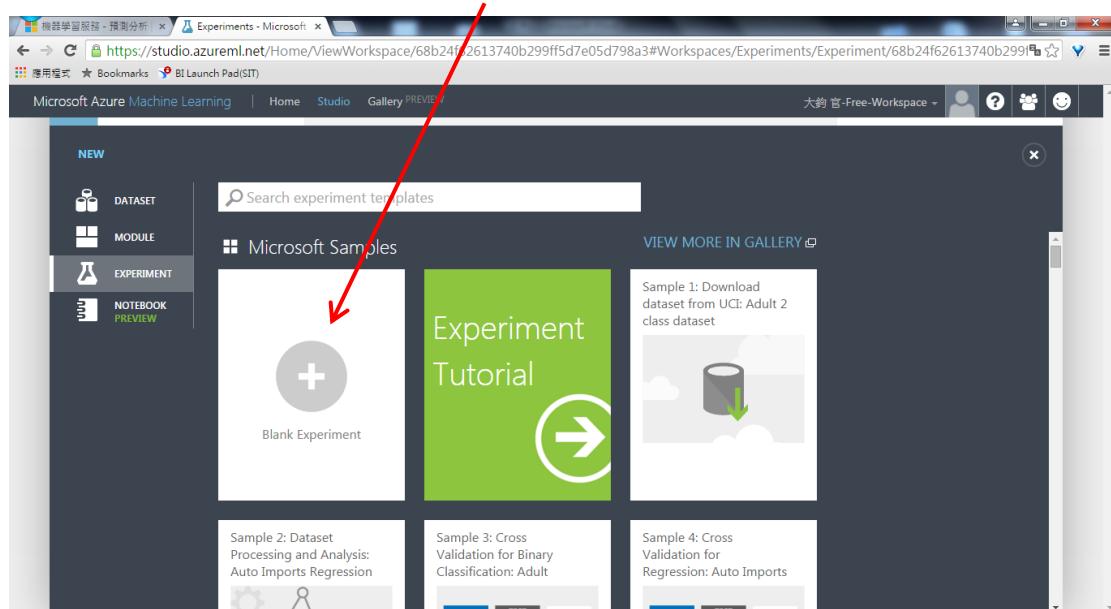


或是依照本 Demo 選擇 Not now，直接進行下一步驟直接使用

4. 點選左下方的 +NEW 新增一個實驗

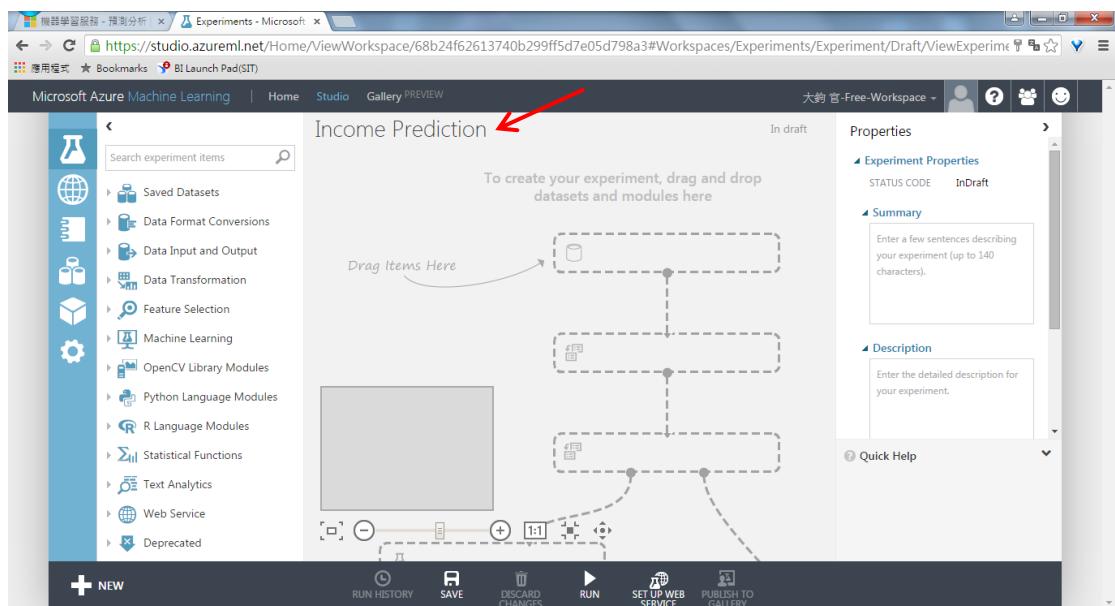
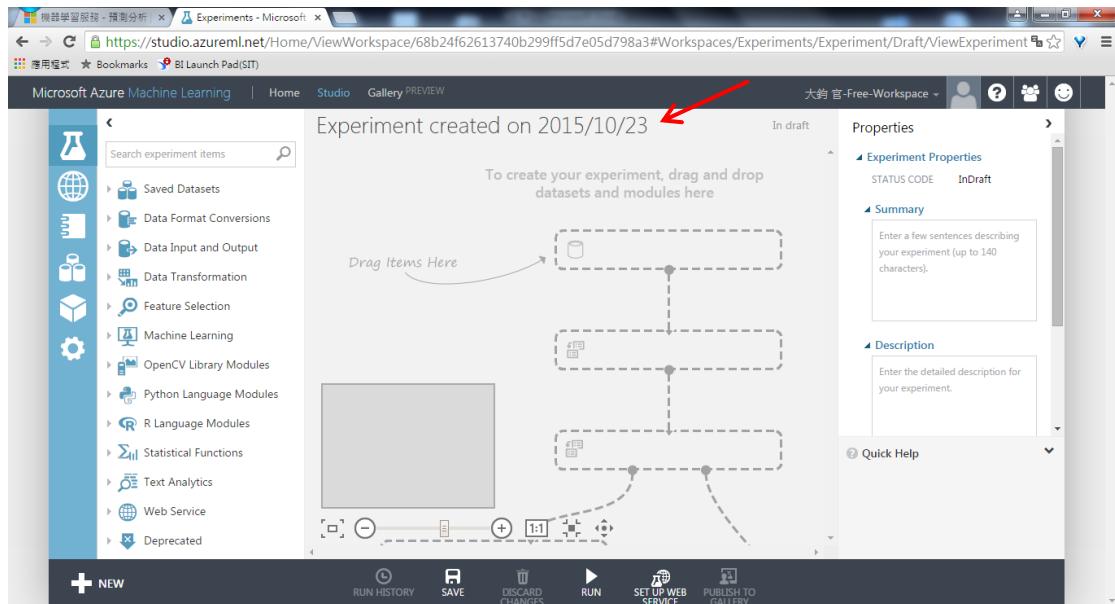


5. 當畫面跳出新的視窗時，點選 Blank Experiment 增加一個實驗

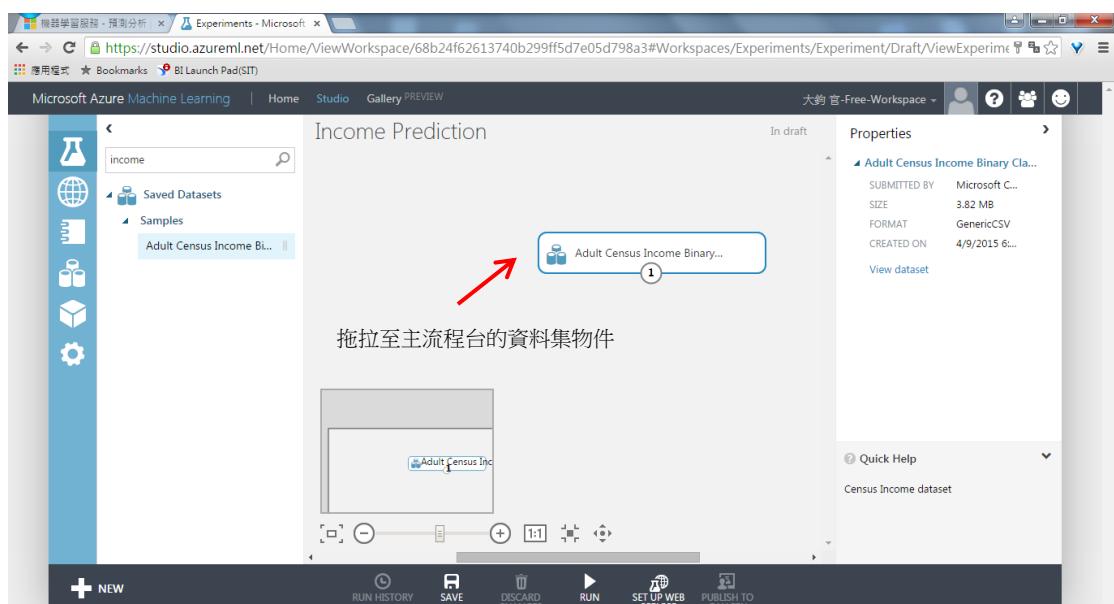
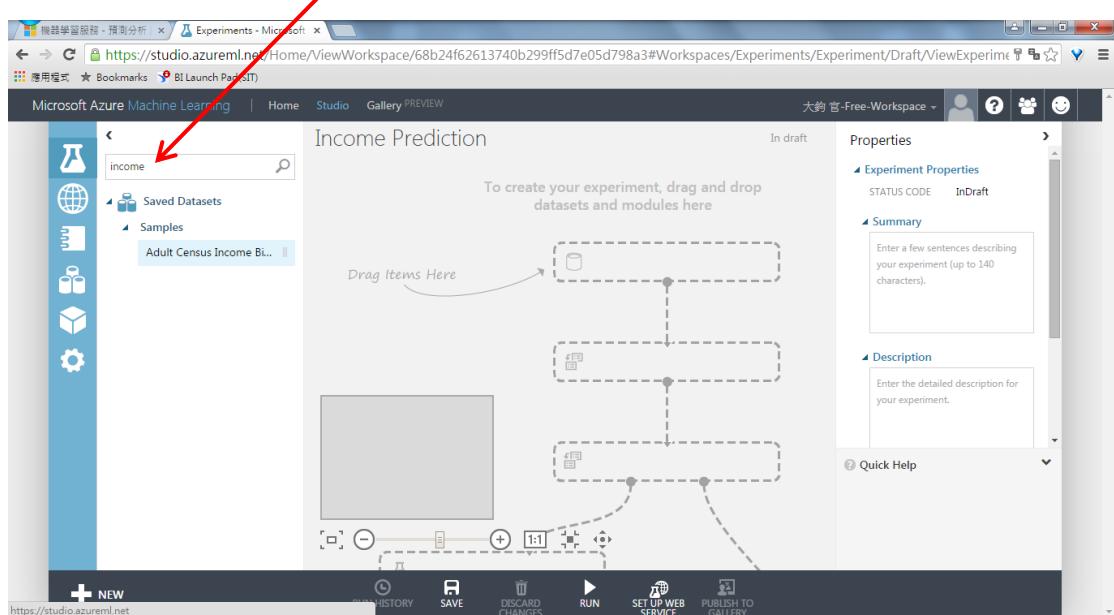


若選擇其他動態磚則可視看一些資料探勘中典型的演算法實作
如本公司會用到的時間序列分析也有出現在其中一個動態磚

6. 畫面將開啟新實驗的頁面，可點選上方更改實驗名稱
此 Demo 將使用範例資料集[Income]做示範，用此資料去做【收入預測分析】
因此我們可以將此實驗名稱訂為【Income Prediction】



7. 左邊的欄位可選擇自己已存入的資料集、資料產出結果、機器學習的演算法、其他程式語言的程式庫、統計方法等…
在此我們採用 Azure 內建的 Income 資料集，所以先於左上方的 Search Bar 中輸入 income，而 Azure ML service 的使用者介面可採拖拉的方式，因此可將搜尋的 income 資料集拖拉到中間的主畫面

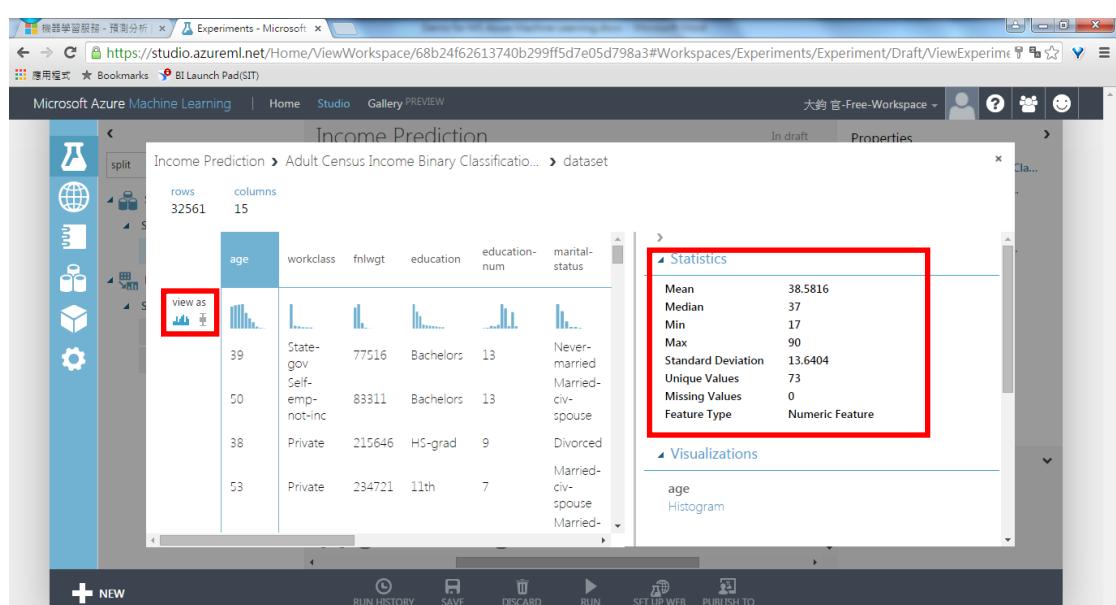
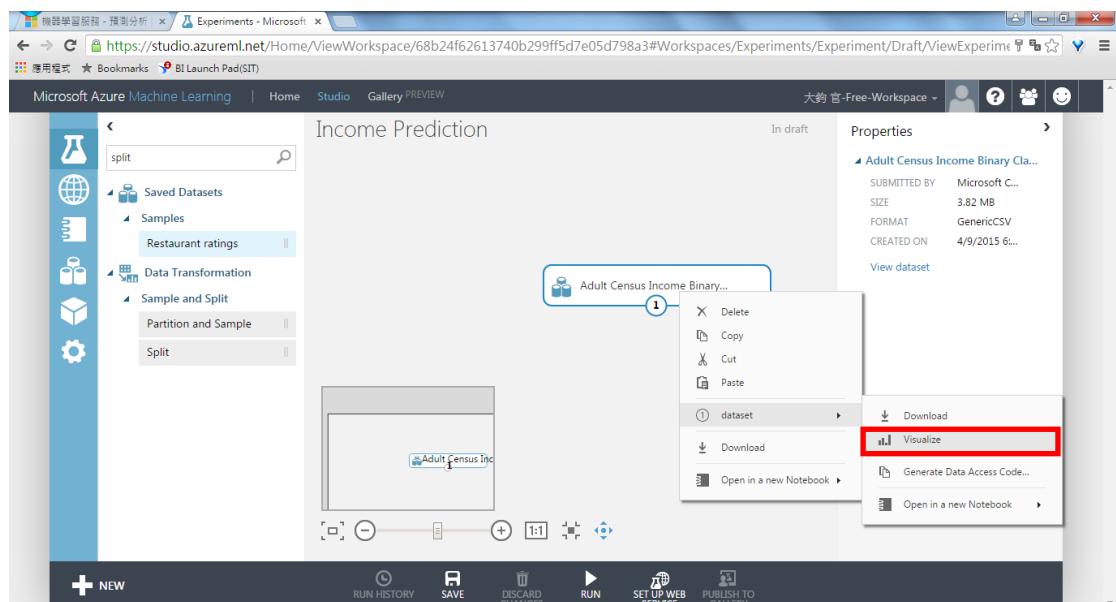


8. 將游標放置於主控台的資料集上，並點選右鍵可選擇 dataset→visualize，則可視覺化觀看資料的主要內容。

新視窗則顯示資料的詳細內容，如各欄位及其記錄，及左上方顯示欄位總數與資料總筆數。

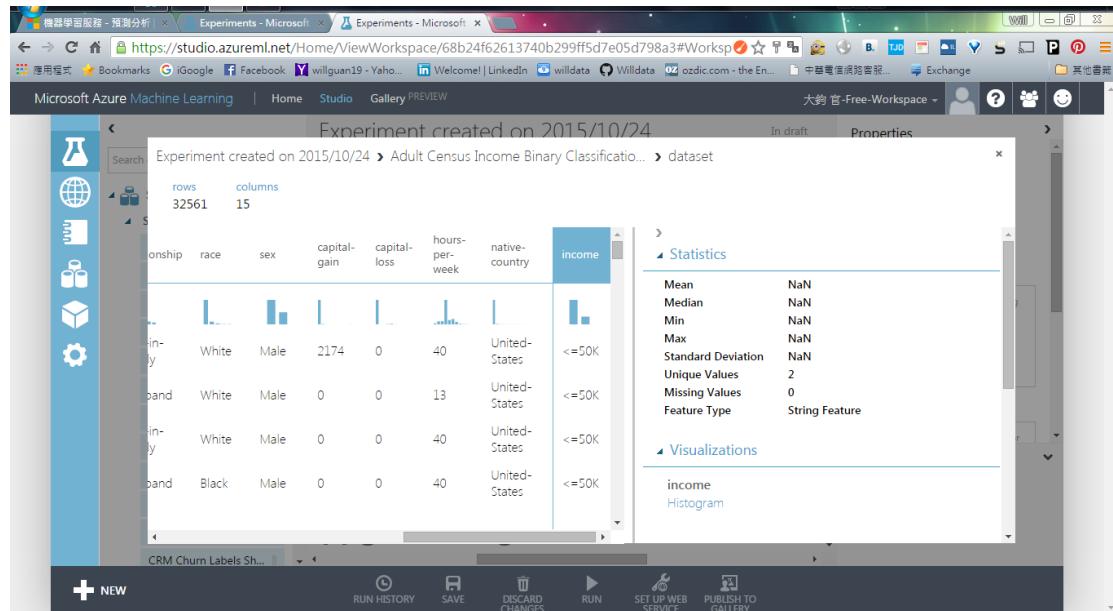
另外，點選欄位時，可於右方觀看該欄位的初步統計資料，如平均數、中位數、最大最小值、標準差和資料型態等…

除此之外，在表格左方的 view as 可依需求選擇資料呈現的圖表，如直方圖或盒鬚圖。

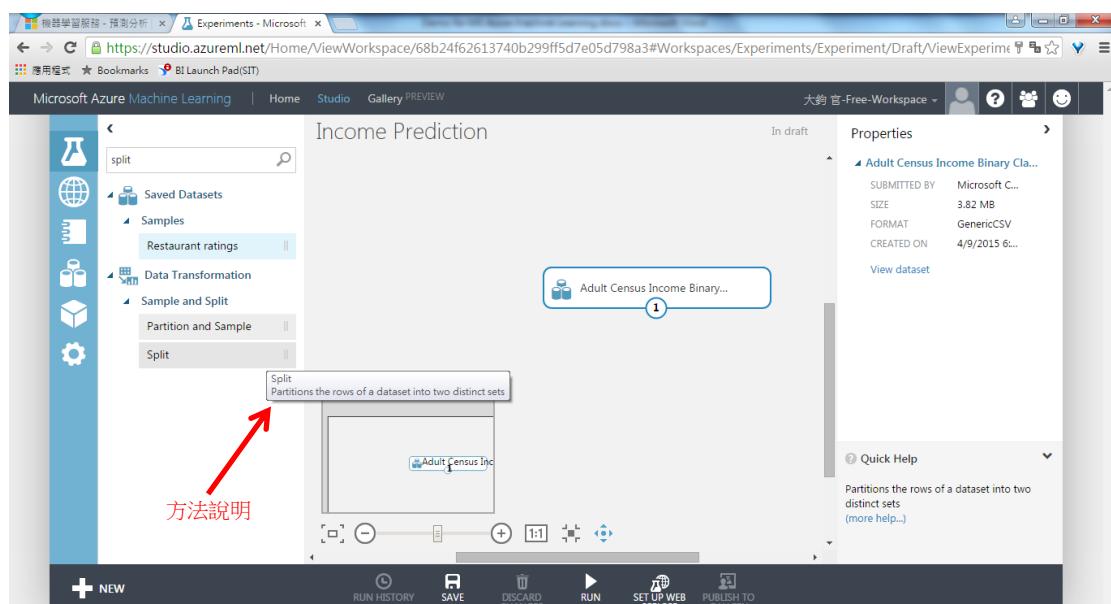
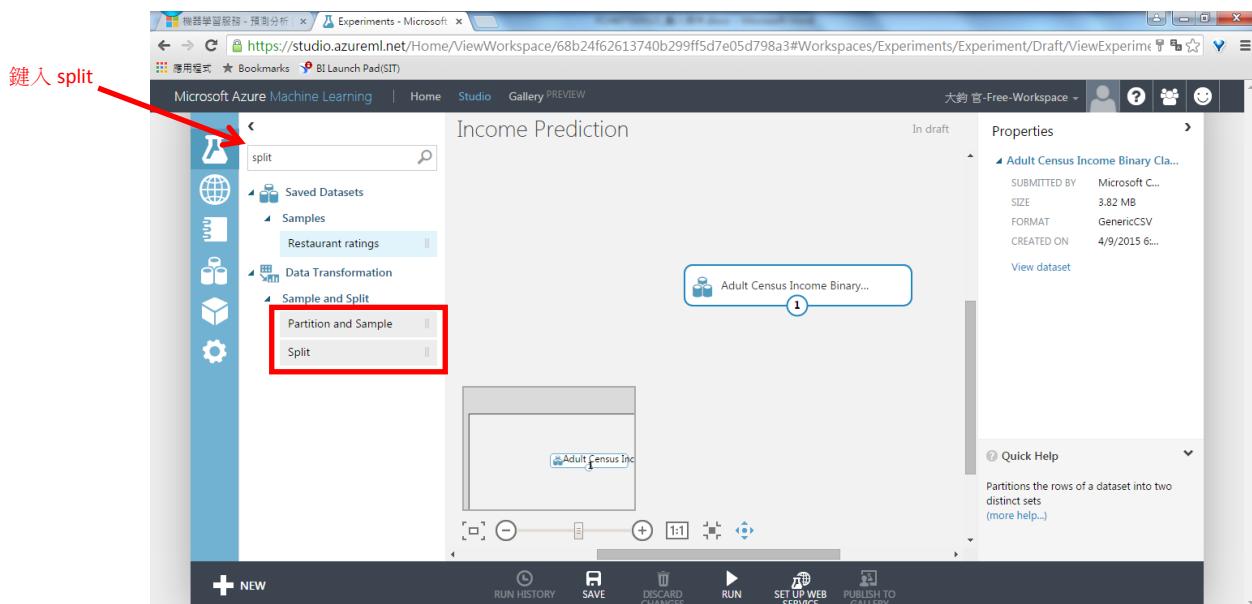


將畫面拉到表格的最右邊，可看到 Income 這個欄位。

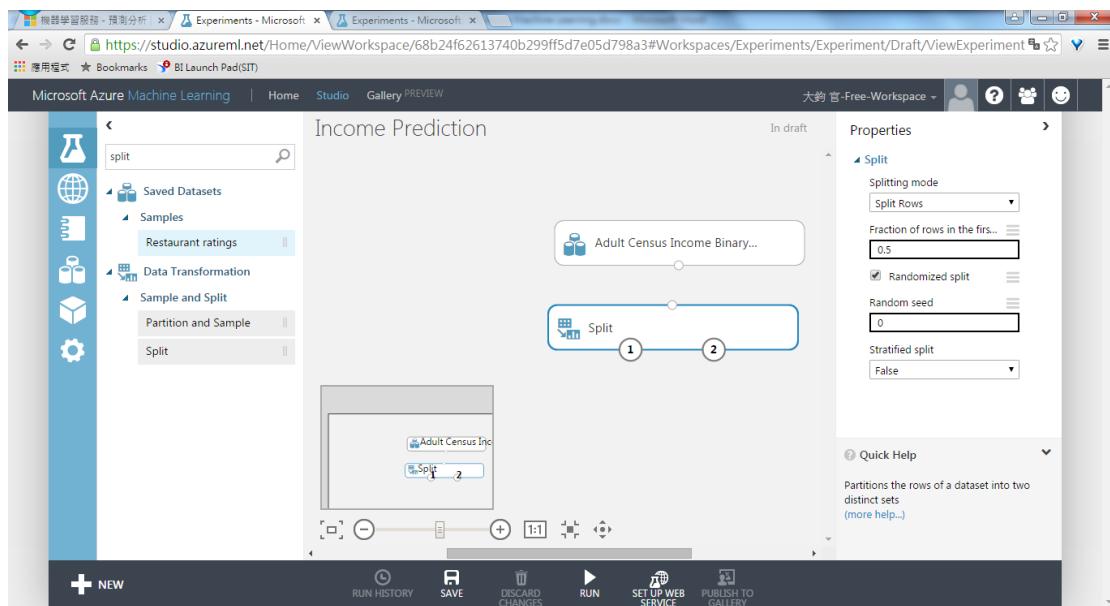
從原始資料中，我們可以得知 Income 這個欄位是採用二分法，收入已 50K 為基準，所以我們之後的預測也是將未來的每筆資料區分為 $\leq 50K$ 或 $> 50K$ ，也就是二分法(Two Class)，待會就會提到關於二分法的演算法。



9. 下一步將資料集區分為 **Training Data** 和 **Test Set**，其原因是因為我們目前沒有未來的資料可測試預測的精準度，所以需要將現有的資料集區為兩部分，一部分是當作訓練模型用的過去資料，另一部分是假裝成未來資料以供我們測試模型的成效。至於要如何區分，一般來說是遵循 **80/20** 法則，即 **80%** 為 **Training Data**，**20%** 為 **Test Set**，也可以另外依資料集的數量狀況斟酌調整比例，如 **70/30** 或 **90/10**。若有資料有時間性，則區分方法須依照時間切割。至於如何切割資料，只需在 **Search Bar** 上鍵入 **split**，即出現兩種資料集切割的方法，將游標移至該方法上時，畫面會介紹該方法的原理。

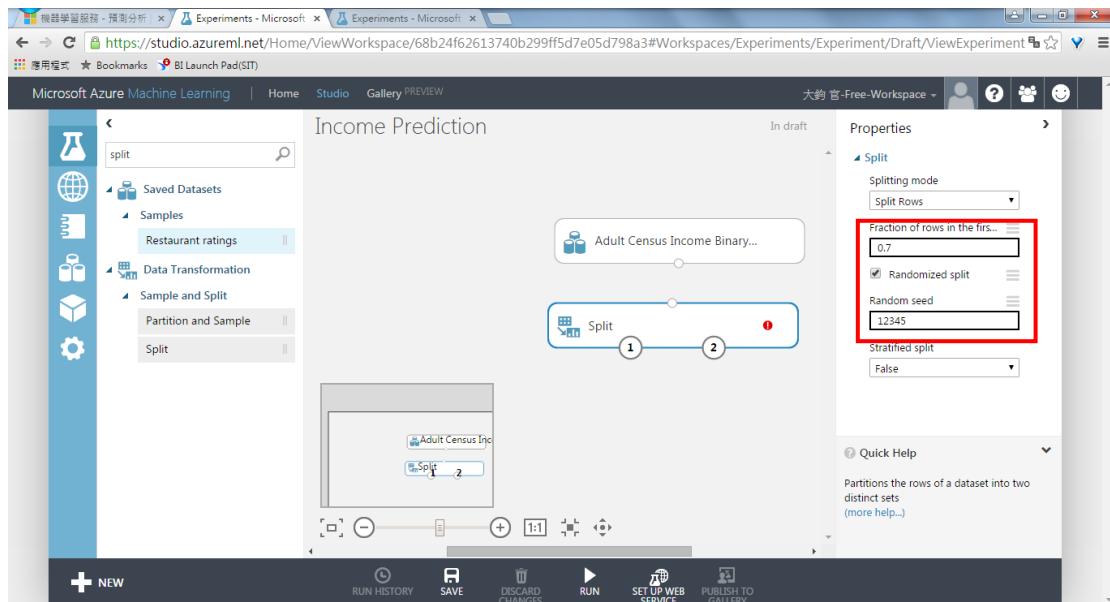


10. 我們在此採用第二種方法，並同樣使用拖拉的方式將切割方法移動到主控台

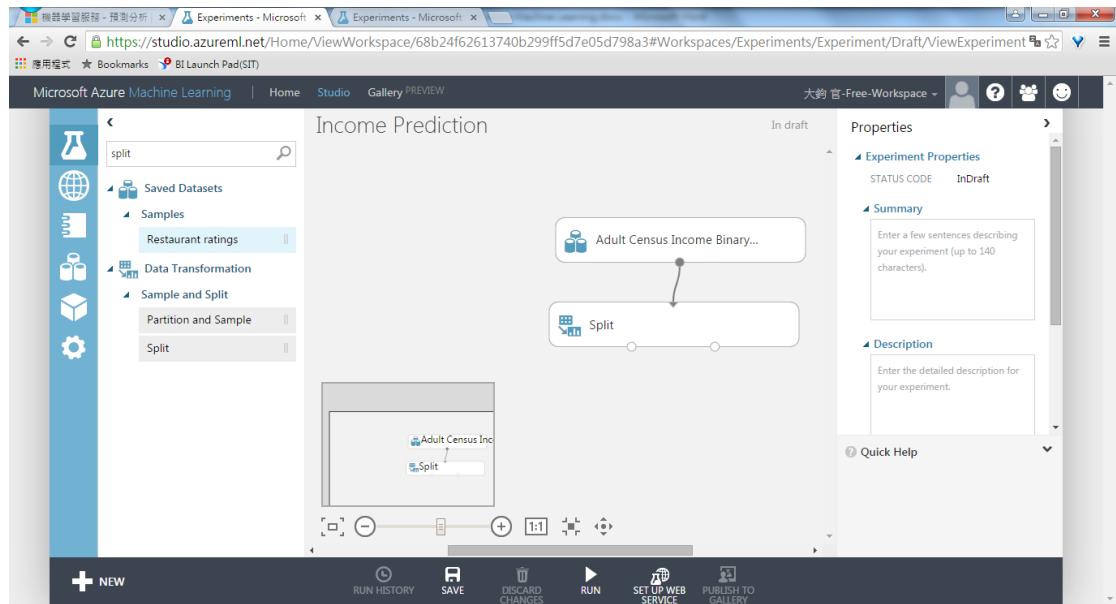


將 Fraction 改為 0.7 (亦即採用 70/30 法則，若資料總數夠大，可採用 90/10 法則將模型訓練至更完善，若資料量不大，則不建議)

Random seed 則隨意填入 12345 即可，亦即隨機數的出發點



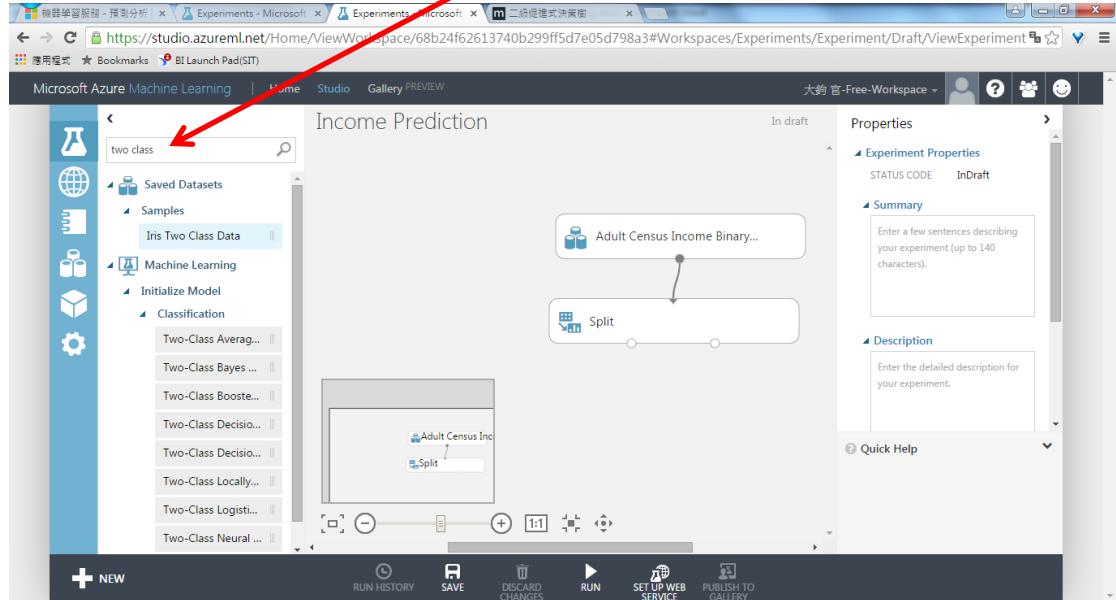
將游標移至資料集處下方的圓圈，並按住滑鼠左鍵，Split 方法的邊框則立即變為虛線，此時按住左鍵不放將游標移至 Split 方框中，放開滑鼠後即出現兩方框的連結線，表示已連結資料集與切割方法。



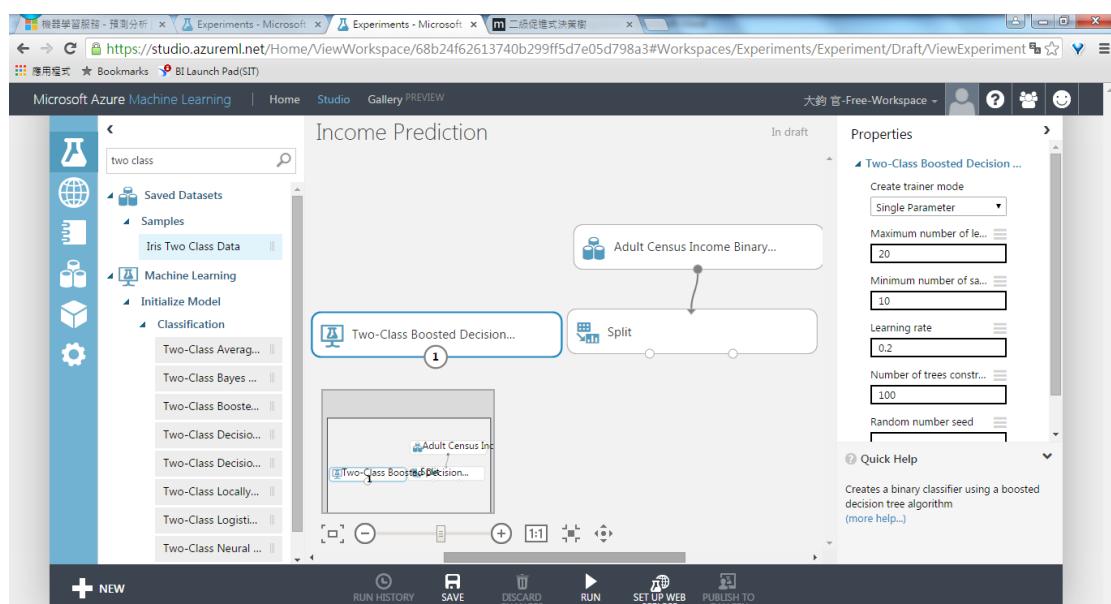
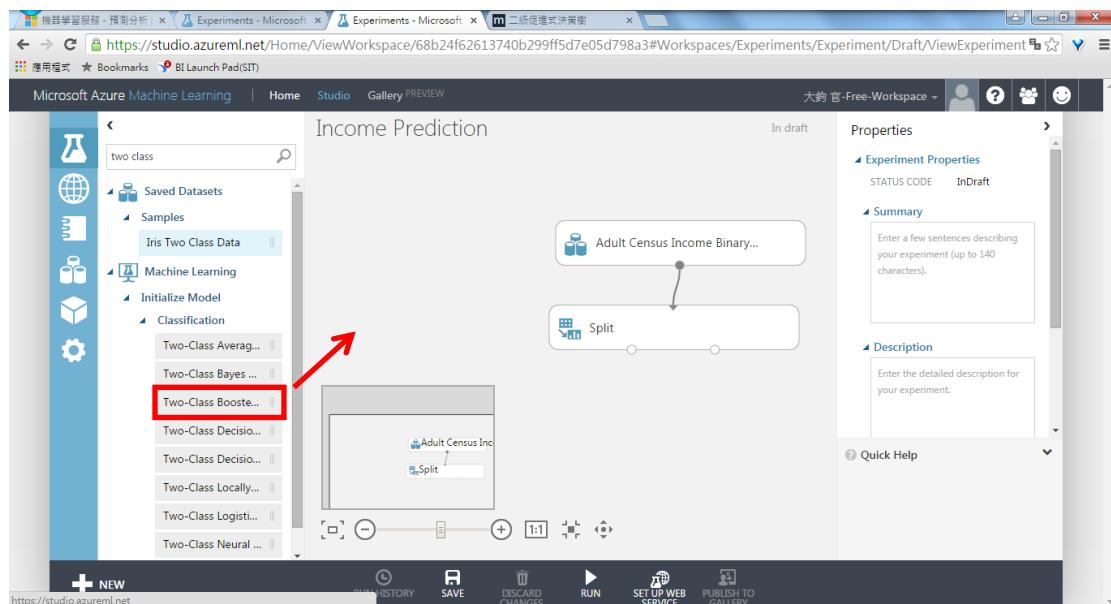
11. 下一步我們將選擇機器學習的演算法，這邊我們採用二級促進式決策樹 (Two-Class Boosted Decision Tree)，以下連結可查閱此決策樹的說明

<https://msdn.microsoft.com/zh-tw/library/azure/dn906025.aspx>

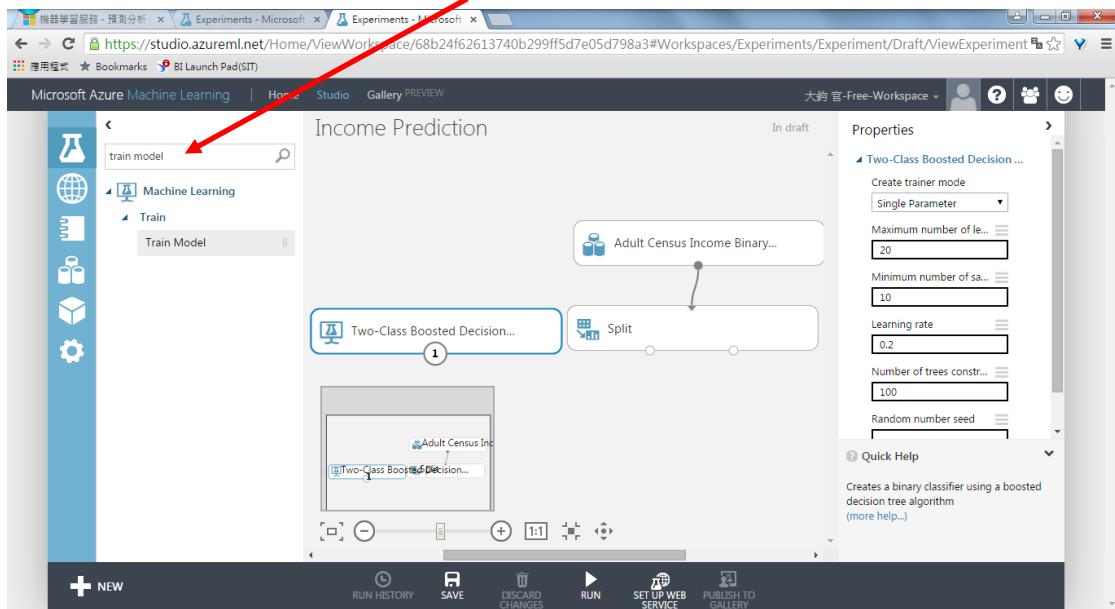
所以我們先於 Search Bar 鍵入關鍵字【two class】，即可出現機器學習中與 two class 相關的分類器



選擇 Two-Class Boosted Decision Tree 分類器，並拖移至主畫面。



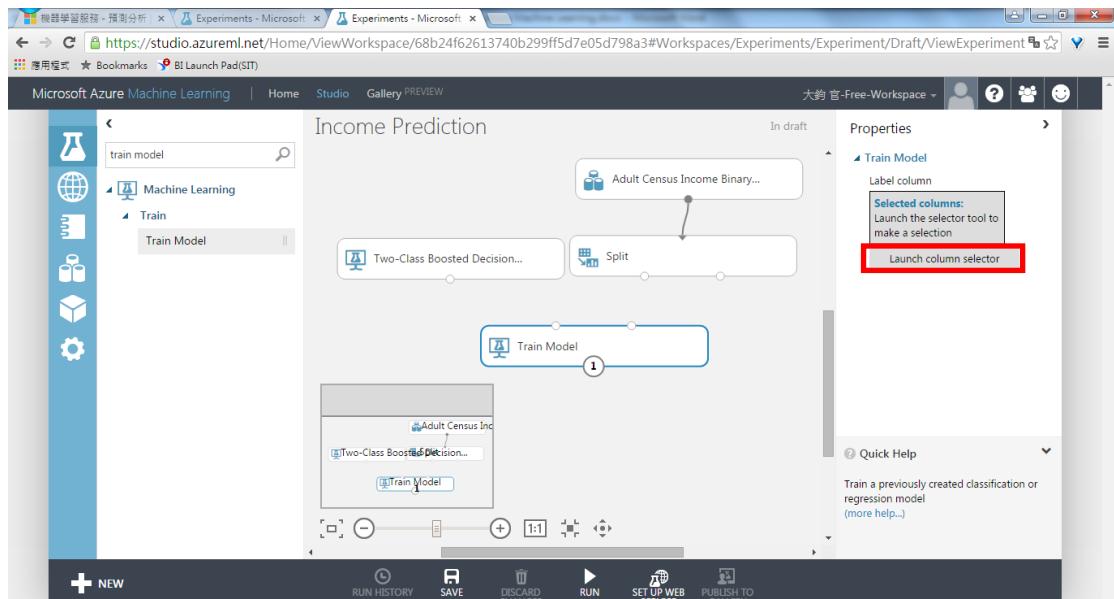
12. 接下來放入訓練模型。先搜尋 train model



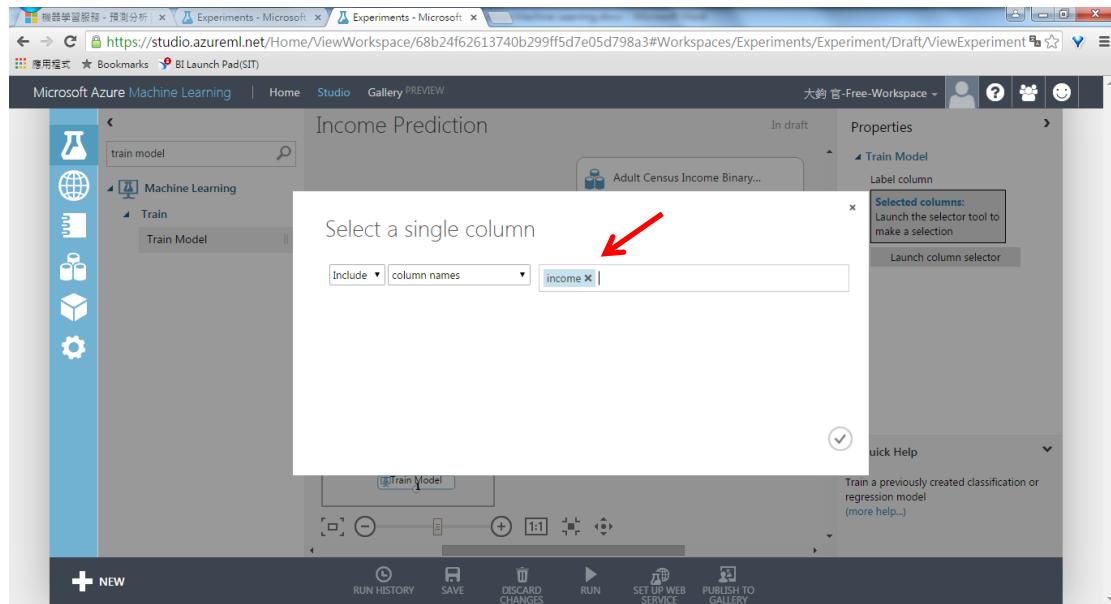
至於為何要放入 Train Model 這個物件，以及為何不是將已切割的資料集和分類方法結合就好。

因為這個 Service 裡的所有東西都是物件，而我們可以看到物件與物件結合是有方向性的(也就是有個箭頭)。若我們將 Two-Class Boosted Decision Tree 與 Split 結合的話，試問是將方法應用到資料集，還是將資料集採取該方法實作？不知道。所以我們採取的策略是在中間放置一個空機器物件，將方法安裝到機器上，也將資料集載入到機器中，接下來就讓機器自己運轉，機器會使用自身擁有的功能(及方法)去運算自身擁有的資料(及資料集)，最後即可產生一個樣本機器。

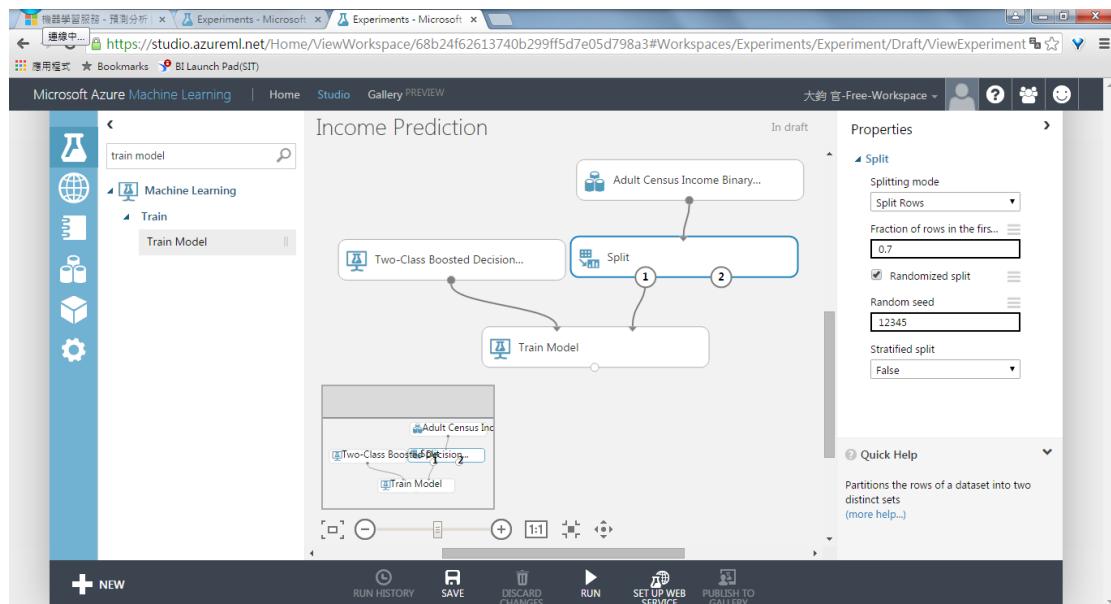
下一步將左方的 Train Model 拖移至主畫面，並點選右方的 Launch column selector



跳出 Launch column selector 小視窗後，我們可以選擇依何種欄位選取方式來選擇欲預測的欄位，此範例中我們依 column names 方式，並鍵入我們要預測的 income 欄位。



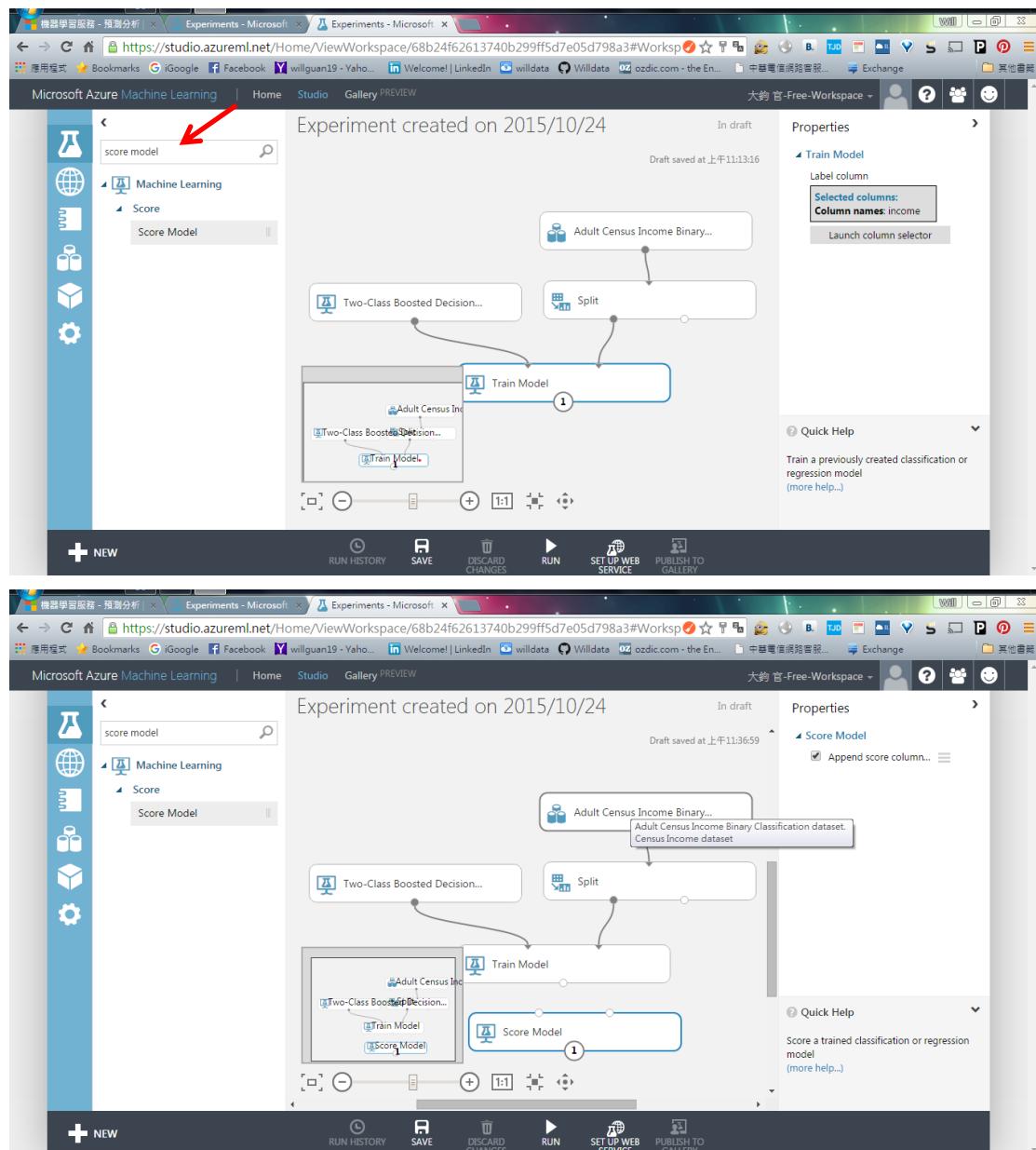
點選右下方的勾勾確認後，將分類器(此即 Two-Class Boosted Decision Tree)與 Train Model 連結，並將切割後的資料集(即 Split)與 Train Model 連結，需要注意的是，Split 需選擇 1 號小圈圈與 Train Model 連結，因為 1 號小圈圈是代表 70%的 Training Data，而我們的 Train Model 須由此 Training Data 來訓練。



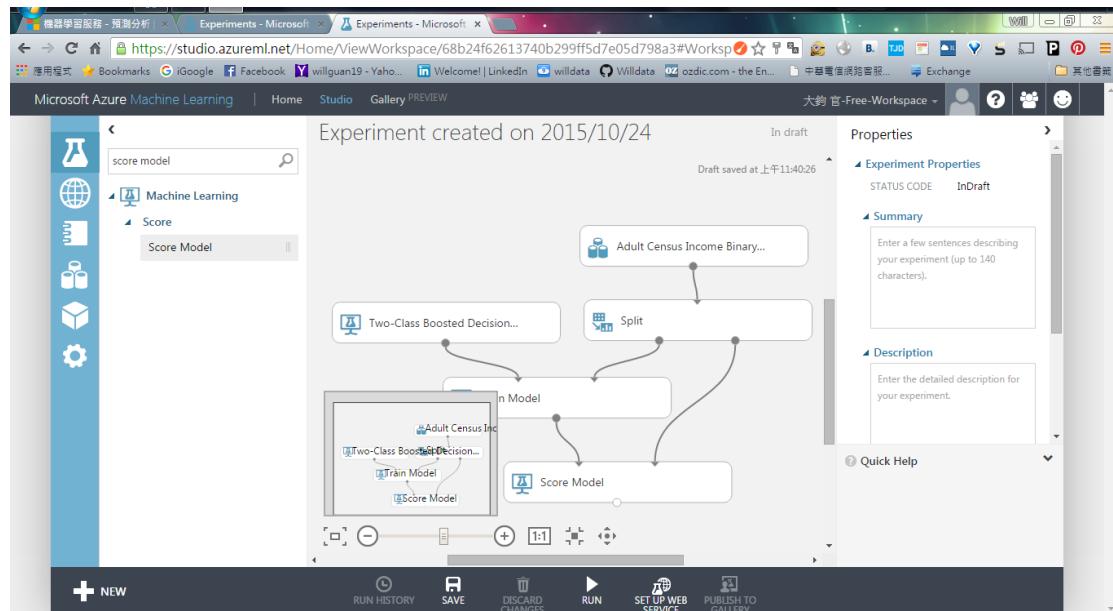
13. 我們將 70%的資料拿來訓練模型後，前置工作都就緒了，接下來就是開始準備用這個模型來預測剩餘的 30%資料。

但是預測這件事沒辦法一步登天，我們需要兩個步驟。第一步，我們需要將過去的經驗(即 Train Model)與未來的狀況(即 30%切割資料)做結合，也就是將 30%的資料一筆一筆都放入決策樹的頂端去跑分類流程，得到我們要的預測結果。第二步，我們確認我們預測結果的好壞，所以我們要在去做一些更細部的分析，去評比這道菜夠不夠格端上桌。

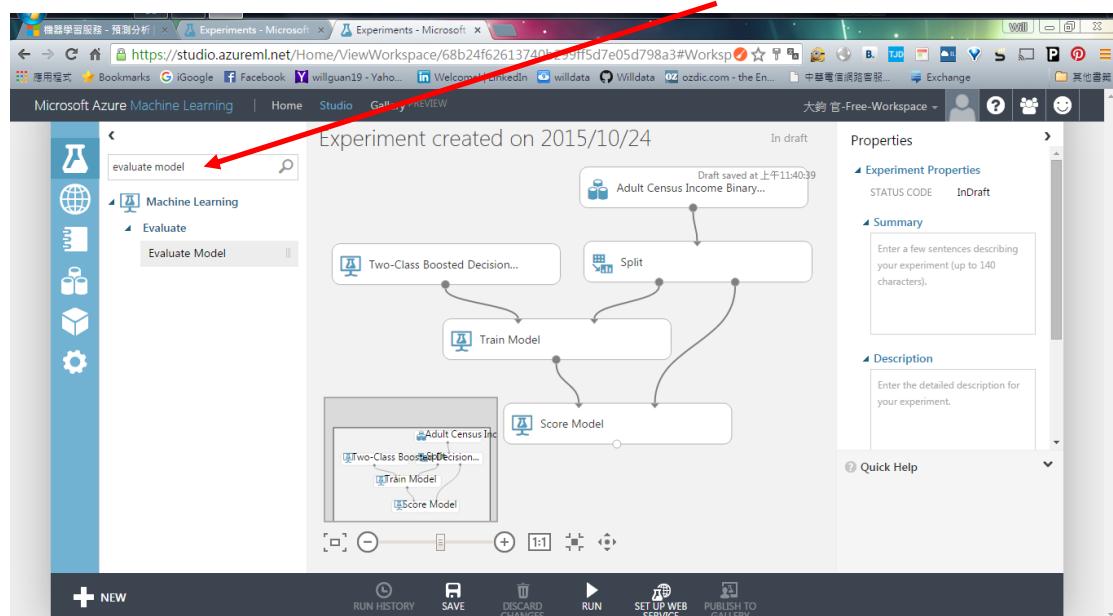
第一步，輸入 score model 這個關鍵字，並將 score model 這個媒介放入主控台來結合經驗與資料



需注意的是，我們要將經驗(Train Model)與待預測的資料(30%Split，也就是Split方框中的2號小圈圈)做結合

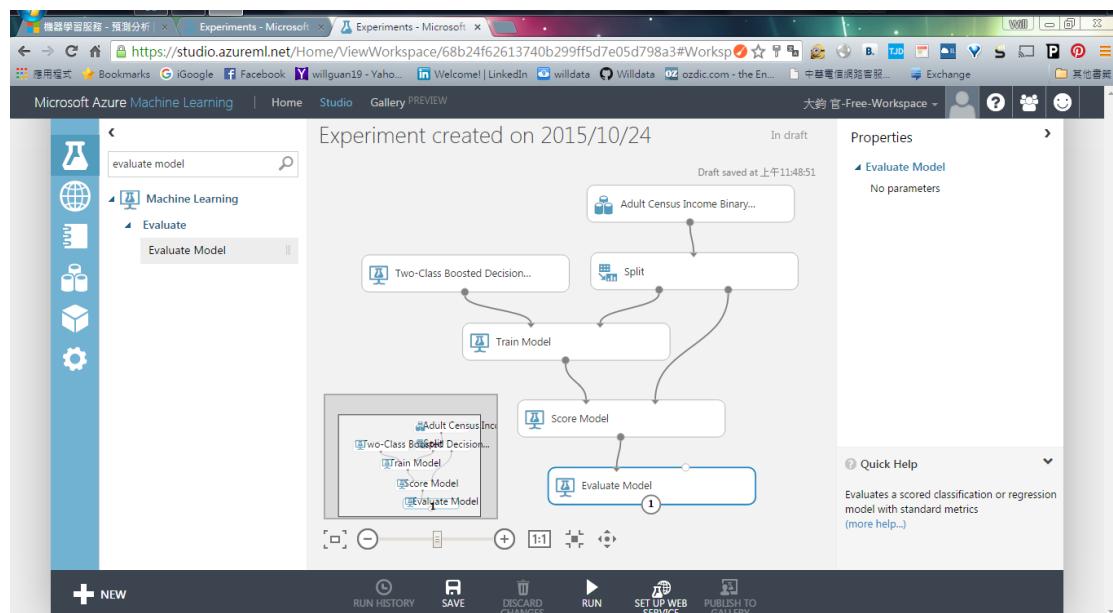


下一步，呼叫專門做細部分析計算的苦力，鍵入 evaluate model

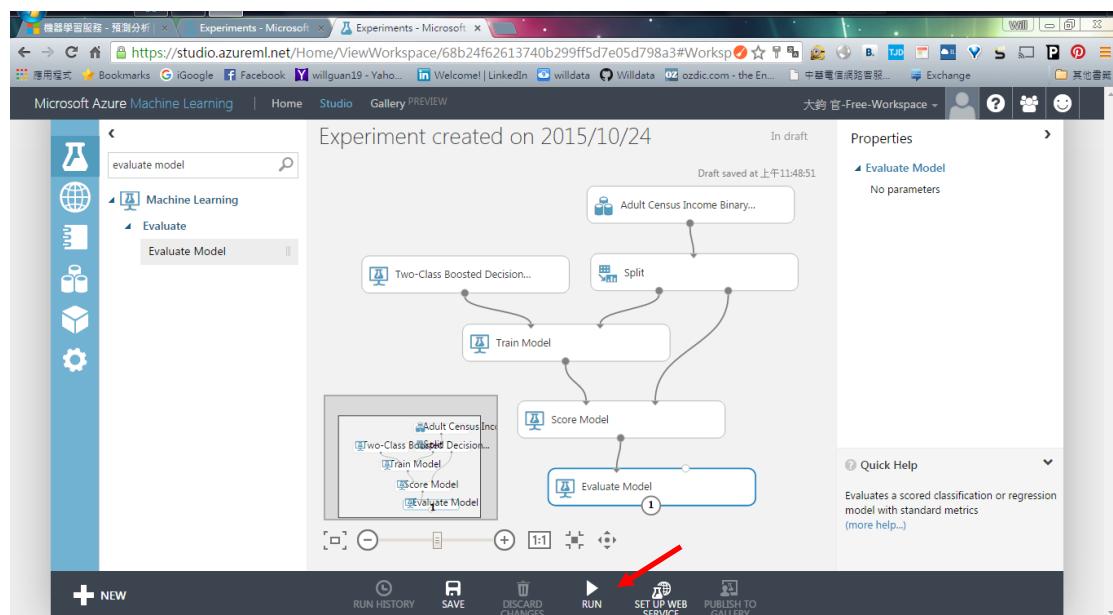


連接 Score Model 和 Evaluate Model，開始請 Evaluate 做苦力。

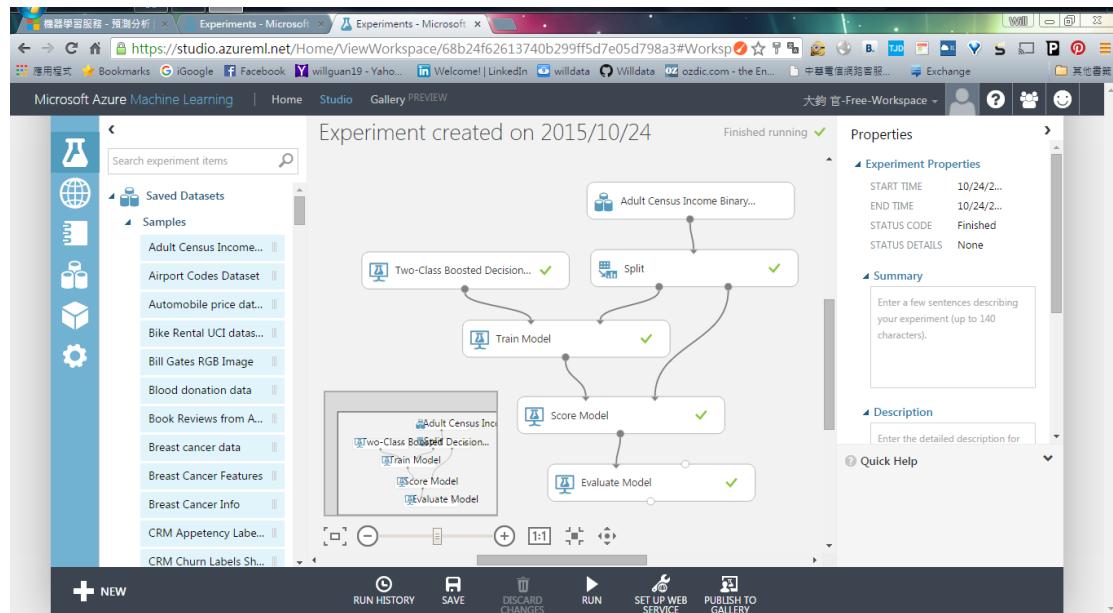
需注意的是，請將 Score Model 與 Evaluate Model 的左上方小圈圈連接，因為右上方小圈圈是用來做模型比較，這個範例中我們不會使用到。



14. 一切準備都準就緒了，點下畫面下方的 RUN，讓系統專心跑整個流程

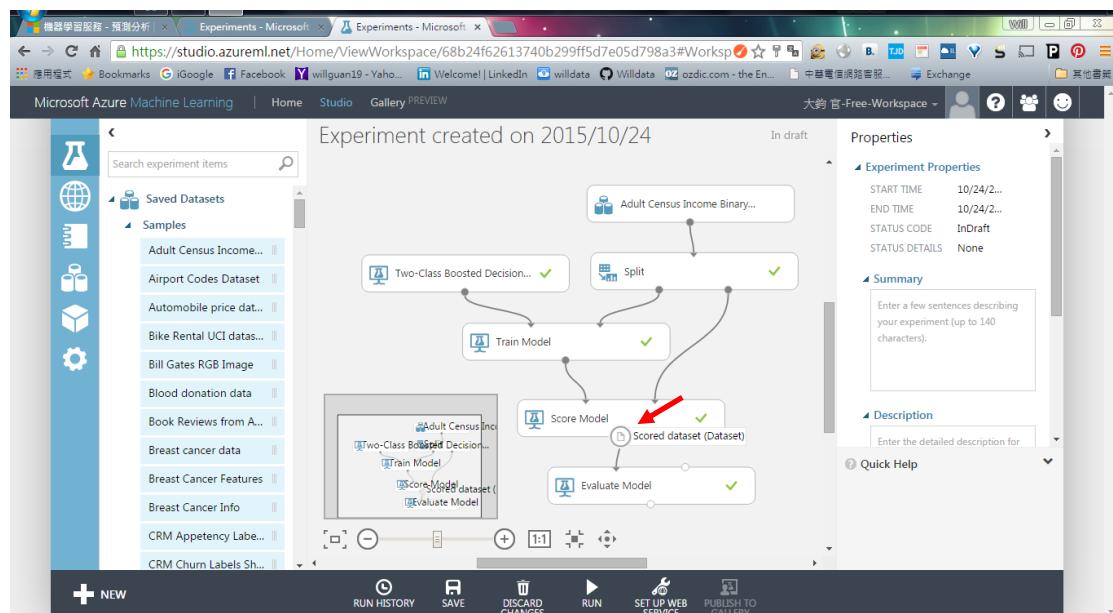


因為資料量不大，所以這塊小蛋糕 Azure 力馬吃完。
看到所有框框都出現綠色勾勾時，表示所有流程與計算都結束了。

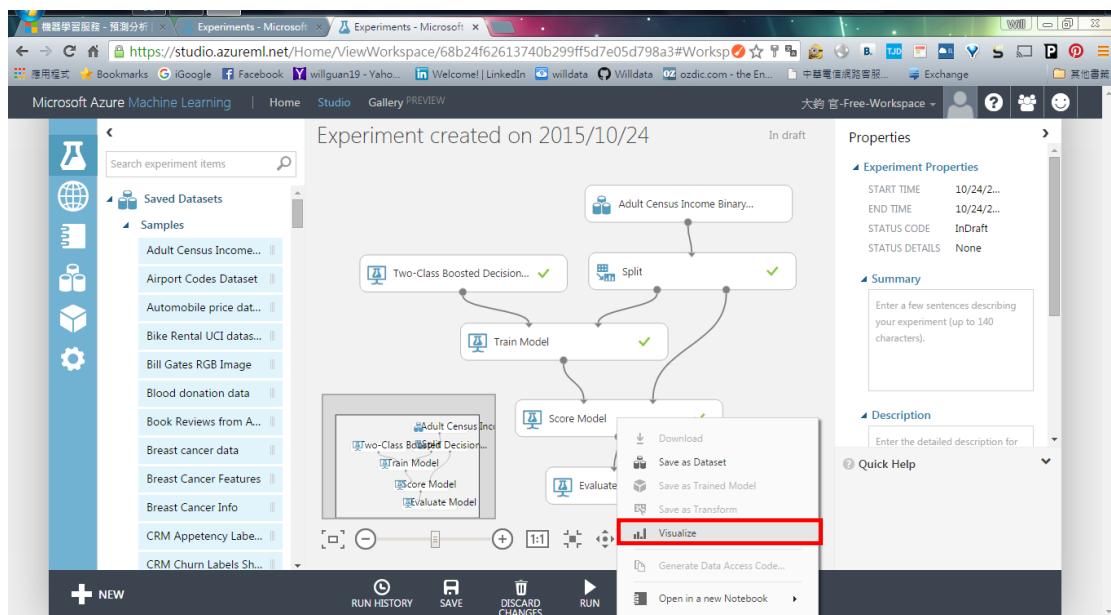


15. 最後一步，我們不要忘了我們的初衷，我們是要預測 Income 的，所以我們要來看看我們預測的情形如何。

首先，將游標移至 Score Model 方框下的小圈圈



按一下左鍵，並點選 Visualize

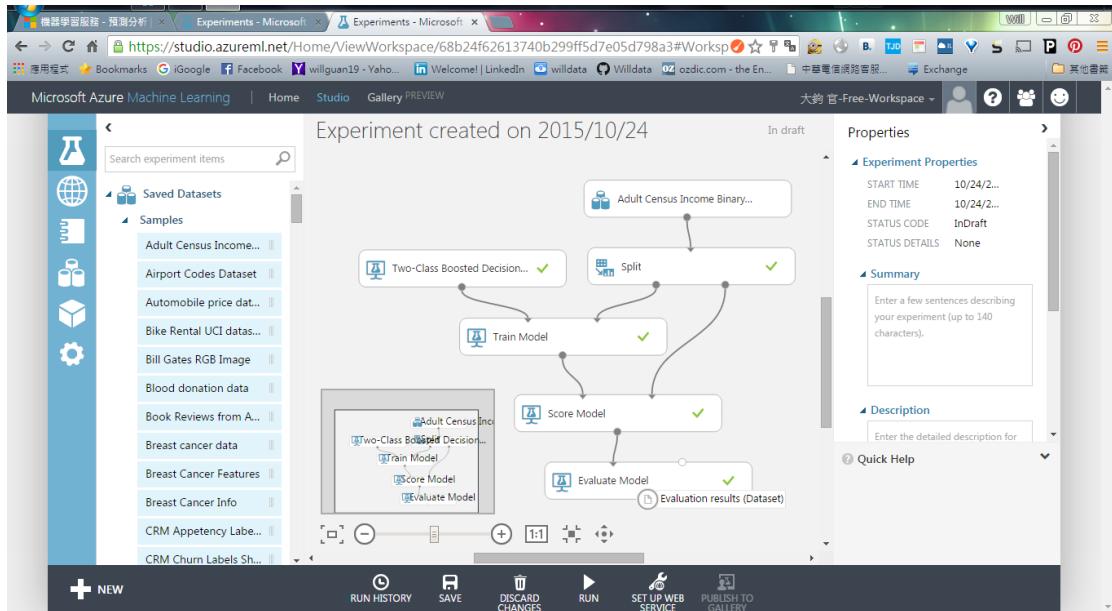


表格的最後兩個欄位則是我們的預測結果

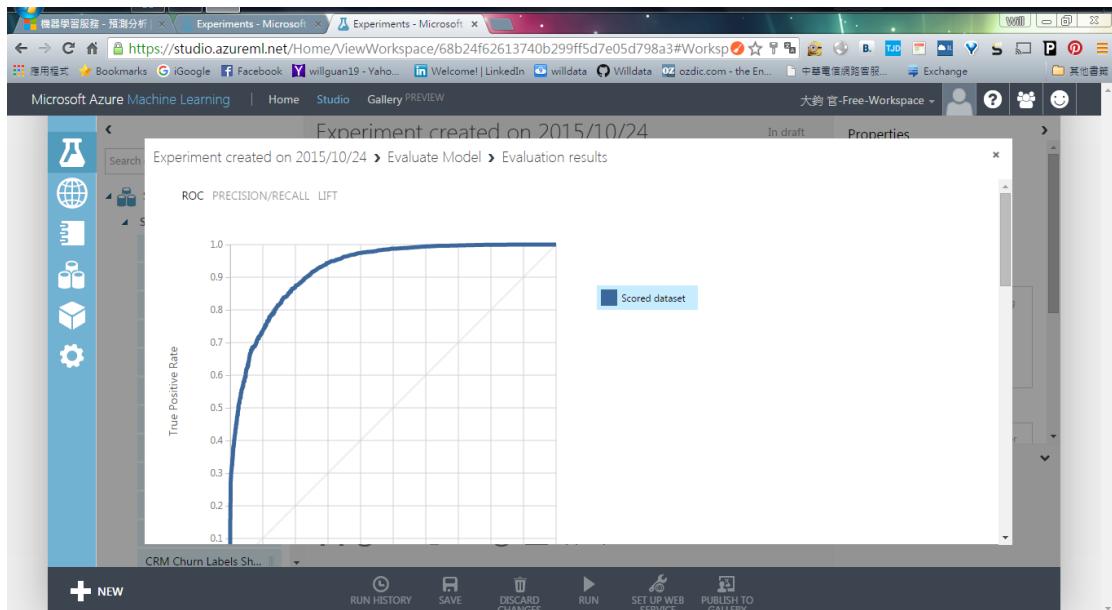
The screenshot shows the 'Scored dataset' view in the Microsoft Azure Machine Learning Studio. The table has 17 columns: capital-gain, capital-loss, hours-per-week, native-country, income, Scored Labels, and Scored Probabilities. The first four columns are categorical, while the last three are numerical. The 'Scored Labels' column contains values like '<=50K' and '0.38015'. The 'Scored Probabilities' column contains values like '0.120344' and '0.053951'. To the right of the table, there's a 'Statistics' section with various metrics like Mean, Median, Min, Max, Standard Deviation, Unique Values, Missing Values, and Feature Type. Below that is a 'Visualizations' section with a 'Scored Labels Histogram' link. The top right panel shows 'Experiment Properties' with details like START TIME: 10/24/2..., END TIME: 10/24/2..., STATUS CODE: InDraft, and STATUS DETAILS: None. The bottom right panel has sections for 'Summary' and 'Description'.

capital-gain	capital-loss	hours-per-week	native-country	income	Scored Labels	Scored Probabilities
0	0	49	United-States	<=50K	<=50K	0.38015
0	0	40	United-States	<=50K	<=50K	0.120344
0	0	40	United-States	<=50K	<=50K	0.053951
0	0	50	Yugoslavia	<=50K	<=50K	0.164289

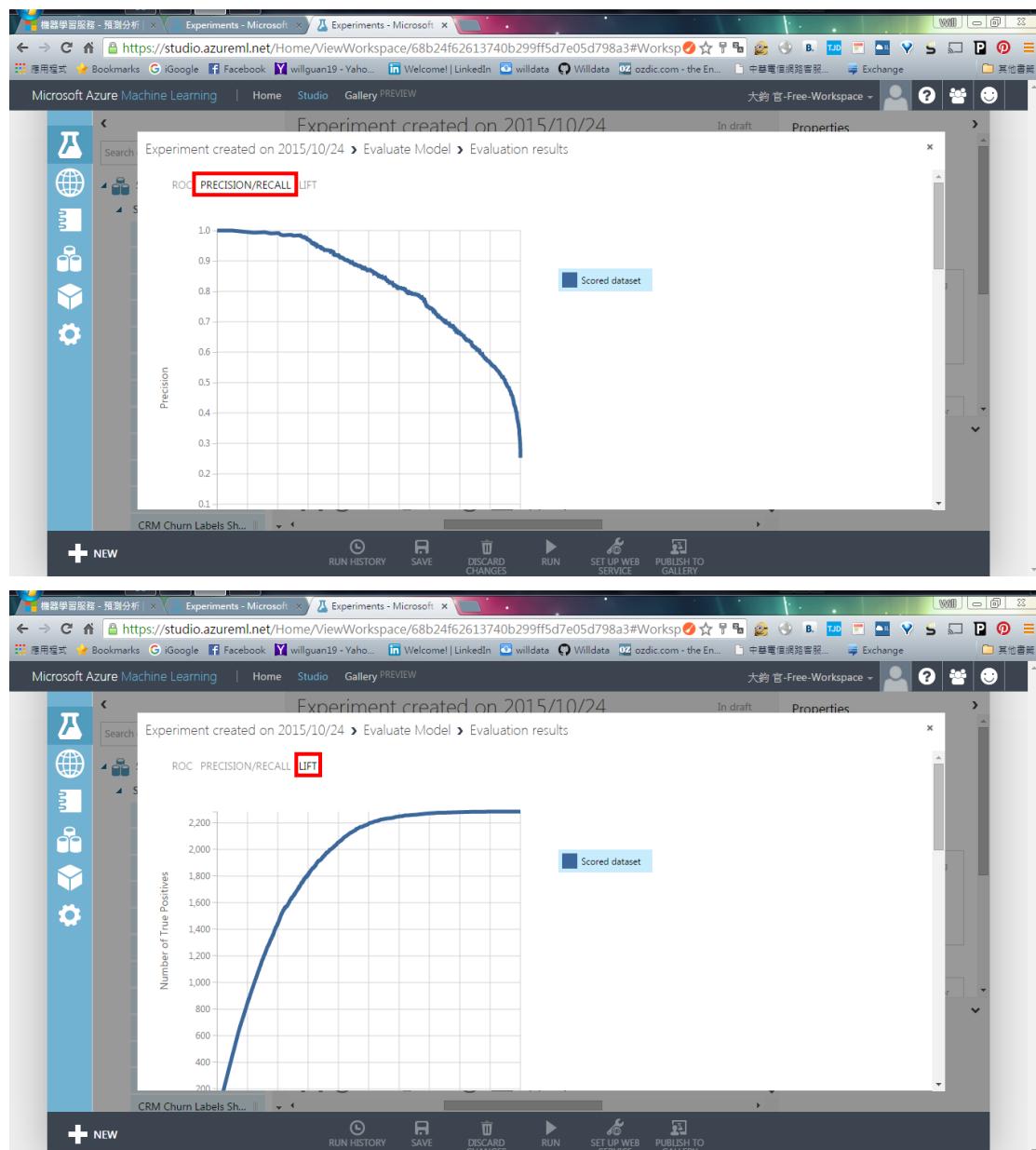
16. 關閉小視窗，將游標移到 Evaluate Model 的小圓圈，同樣點選 Visualize 檢視細部分析的結果。



我們可以看 ROC 曲線(**receiver operating characteristic curve**)來驗證我們的模型是否可靠。若曲線是在明顯在對角線上方的話，表示這是個好模型；若是在對角線下方，或是有點模凌兩可，不上不下的話，表示還有待改進。



當然，我們也可以點選 PRECISION/RECALL 或 LIFT 曲線來檢視我們的模型。

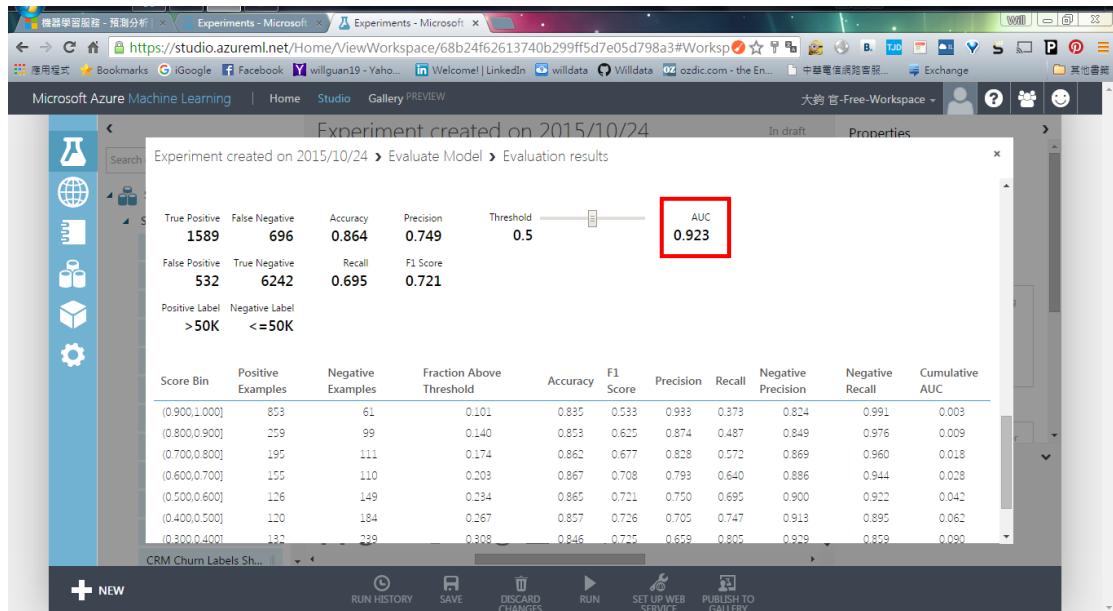


我們可以將畫面拉到下面，看更多的數據。

如果要詳細了解各數據的意義，請看 ROC 曲線的說明

https://zh.wikipedia.org/wiki/ROC_曲線

以 ROC 曲線來看的話，主要以 AUC 數值為準，0.5 等於是擲銅板亂猜；小於 0.5 就表示比銅板還不如，用這種模型不如請周星馳擲銅板，除非把這模型當作反指標來看，那就有其意義；大於 0.5 就表示正確率很高，是個好模型。



以上是基本的實驗操作，此操作流程與微軟的流程示範相同。此份教學若有錯誤之處，請不吝告知，勘誤交流信箱：aieren61will@gmail.com

這是數字 123 的 1

This is the lowercase of L, will for Will Smith

若有其他主題感興趣的話，亦可聯絡本人，我會再詳加補充。Thanks!