



# 第一屆 SAS 校園資料採礦競賽

## IMTKU 淡江資管隊

### 專案說明文件

報告組別：IMTKU 淡江資管隊（淡江大學資管所碩士班）  
指導老師：戴敏育 博士 (Dr. Min-Yuh Day)  
隊長：杜駿 (Chun Tu)  
隊員：  
陳維君 (Wei-Chun Chen)  
許安琪 (An-Chi Hsu)  
黃世禎 (Shih-Chen Huang)

2012 年 09 月 21 日

## 目錄

一、 緒論 .....	1
1.1 專案研究背景 .....	1
1.2 專案研究動機 .....	1
1.3 專案研究目的 .....	1
二、 文獻探討 .....	2
2.1 決策樹(Decision Tree) .....	2
2.2 類神經網路(Neural Network) .....	2
2.3 線性迴歸模式(Linear Regression) .....	3
2.4 支撐向量機(Support Vector Machine) .....	3
三、 專案流程 .....	4
3.1 應用 SAS Enterprise Guide 整合資料及變數延伸 .....	4
3.1.1 原始資料 .....	4
3.1.2 延伸變數-行內行為資料 .....	6
3.1.3 延伸變數-行外行為資料 .....	8
3.1.4 延伸變數-產品持有資料 .....	10
3.2 應用 SAS Enterprise Miner 建立預測模型 .....	14
3.2.1 輸入資料 .....	14
3.2.2 資料切割 .....	22
3.2.3 模型建立 .....	23
3.2.4 模型比較 .....	30
3.2.5 評分資料 .....	32
3.2.6 匯出 .....	35
四、 研究結果 .....	38
4.1 模型過度配置的觀察與處理 .....	38
4.2 資料集與模型比較與結果 .....	38
4.3 結果驗證 .....	40
4.4 結論 .....	42
五、 參考文獻 .....	42

## A1: 附錄一 SAS Enterprise Guide (EG)操作流程 Step by Step 畫面

A1-1 SCHOOL_CCD 和 SCHOOL_ULC 之欄位新建擴增 .....	A1-1
A1-2 SCHOOL_CCDN、SCHOOL_ULCN、SCHOOL_CUSTOMER 和 SCHOOL_TARGE 做連結.....	A1-4
A1-3 SCHOOL_PRODUCT_HOKDING 之欄位新建擴增 .....	A1-9
A1-4 SCHOOL_CUS_CCDULC_TAR、SCHOOL_PRO24N 及 SCHOOL_PRO24B 做連結 .....	A1-14

## A2: 附錄二 SAS Enterprise Miner 操作流程 Step by Step 畫面

A2-1 模組建置前置作業 .....	A2-1
1.1 新建專案(New Project).....	A2-1
1.2 新建類別庫(New Library) .....	A2-2
1.3 新建資料來源(Create Data Source).....	A2-4
1.4 新建圖表(New Diagram) .....	A2-10
A2-2 訓練模組建置 .....	A2-11
2.1 Replacement 節點 .....	A2-12
2.2 資料分割(Data Partition).....	A2-13
2.3 決策樹(Decision Tree).....	A2-14
2.3.1 自動建置樹.....	A2-14
2.3.2 手動建置樹.....	A2-16
2.4 Impute 節點 .....	A2-20
2.5 迴歸(Regression)模型 .....	A2-21
2.6 類神經網路(Neural Network)模型 .....	A2-23
2.7 SVM 模型 .....	A2-24
2.8 模型比較(Model Comparison).....	A2-25
A2-3 評分 .....	A2-26
3.1 評分步驟.....	A2-26
3.2 評分結果.....	A2-28

## 表目錄

表 3.1-1：顧客基本資料，共 21 欄。	5
表 3.1-2：行內行為資料，共 5 欄。	5
表 3.1-3：行外行為資料，共 4 欄。	5
表 3.1-4：產品持有資料，共 4 欄。	6
表 3.1-5：預測目標變數，共 2 欄。	6
表 3.1-6：行內行為資料 Type1	7
表 3.1-7：行內行為資料 Type2	8
表 3.1-8：行外行為資料 Type1	9
表 3.1-9：行外行為資料 Type2	10
表 3.1-10：產品持有資料 Type1	11
表 3.1-11：產品持有資料 Type2	13
表 3.2-1：SAS_0731 資料集	15
表 3.2-2：SAS_0801 資料集	18
表 3.2-3：SAS_0803 資料集	22
表 4.1-1: 三筆資料集之比較	38
表 4.2-1：累積增益數據比較表	39

## 圖目錄

圖 2.4-1 : SVM 超平面分割示意圖.....	4
圖 3.2-1 : 資料切割 .....	22
圖 3.2-2 : Data Partition 參數調整 .....	23
圖 3.2-3 : 自建樹執行.....	23
圖 3.2-4 : 自建樹輸出結果.....	24
圖 3.2-5 : 手動樹執行.....	24
圖 3.2-6 : 手動樹節點分割.....	25
圖 3.2-7 : 插入 Impute 節點 .....	25
圖 3.2-8 : Impute 參數調整 .....	26
圖 3.2-9: 迴歸參數使用逐步回歸 .....	26
圖 3.2-10: 迴歸參數調整.....	27
圖 3.2-11: 迴歸輸出結果。 .....	27
圖 3.2-12 : 插入類神經節點.....	28
圖 3.2-13: 類神經參數調整.....	28
圖 3.2-14: 類神經輸出結果.....	29
圖 3.2-15 : 插入 SVM 節點 .....	29
圖 3.2-16: SVM 參數調整.....	30
圖 3.2-17 : SVM 輸出結果 .....	30
圖 3.2-18: 模型比較 .....	31
圖 3.2-19: ROC 圖 .....	31
圖 3.2-20: 累計增益表.....	32
圖 3.2-21: 模型比較後方，插入 SCORE 節點。 .....	32
圖 3.2-22: SCORE 執行結果.....	33
圖 3.2-23 : 檢視 SCORE 節點執行後結果資訊-點選 EXPORY DATA .....	33
圖 3.2-24 : 檢視 SCORE 節點執行後結果資訊-點選 Browse。 .....	34
圖 3.2-25: SCORE 輸出結果資訊.....	34
圖 3.2-26:觀察 TARGET 為 1 之資訊.....	35
圖 3.2-27:插入 SAS CODE 節點 .....	35
圖 3.2-28:匯出 CODE 撰寫 .....	36
圖 3.2-29:排序之程式碼.....	36
圖 3.2-30:取前 10000 名之程式碼 .....	37
圖 4.2-1 : 各資料表的最佳模型 ROC Chart 結果 .....	39
圖 4.2-2 : 提出 10000 筆建議銷售名單部分截圖 .....	40
圖 4.3-1:原始資料經類神經模型訓練後之預測 TARGET 為 1 之結果資料與經預測後之 5 萬筆評分資料的前一萬筆資料比較.....	41

# 一、緒論

## 1.1 專案研究背景

台灣的金融市場逐漸走向自由化與國際化、金融機構激烈競爭、新興金融商品不斷引進國內市場，其中又以信用卡的快速普及最深入影響大眾生活，國內信用卡市場之發卡量及流通量雖每年不斷在成長，小小的一張塑膠貨幣、而今已經融入民眾日常生活當中、但發卡銀行於設計促銷活動是否有考慮到消費者行為之異質性，以及促銷活動所真正影響之消費族群為何，都是值得深入研究之方向。

一大型金控公司經過幾年的積極擴展爭取新客戶與擴大市場占有率，在信用卡事業處的客戶數有顯著成長，也積極經營產品跨售，銀行管理階層針對信用卡客群，訂定善用信用卡客戶基礎、加強銀行產品交叉銷售，以發揮金控綜效，提升客戶貢獻度為其經營目標。

## 1.2 專案研究動機

根據經濟研究單位對未來兩年的景氣預測發現，基於歐債危機的陰影，對景氣成長持並不普遍看好。因此銀行相對保守，對於高風險的放款產品仍持相對保守的態度。

由於交叉銷售不但能提供顧客「一次購足」的產品與服務之需求，對企業而言，產品的商業價值不止於產品本身的銷售利潤，還有其交叉銷售其他產品的利益。但許多企業因缺乏一套有效的機制，來評估、識別顧客的區隔和屬性，因此在跨類商品的銷售時常無法符合顧客真正的需求，使得銷售績效難以提升。本研究中，採用資料採礦技術建置行銷預測模型，並應用此模型結果針對現行客戶作評分，提出1萬名銷售建議名單，使交叉銷售更有效執行。

## 1.3 專案研究目的

根據上述的研究背景與動機，本研究目的可歸納為以下幾點：

- 一、加強信用卡客戶銀行產品交叉銷售，並考慮外在環境，應用行內風險指標排除高風險客群，加強風險控管。
- 二、在考慮通路執行成本下，提升行銷成交率，增加產品效益。
- 三、如何從交易中瞭解顧客的消費特性進而加強競爭力，同時發覺不同類型的信用卡用戶，進行適當的客戶關係管理，增加客戶的忠誠度及使用率。

## 二、文獻探討

### 2.1 決策樹(Decision Tree)

決策樹是強大且普遍使用的分類或預測工具，它以樹狀的方式由上而下表現出規則，將特定的物件集合，隨著樹的成長，逐步的分割成更小的子集合。決策樹可以產一組規則的集合，可以從新的資料集合產生預測，這個資訊可以幫助企業決策。決策樹超越其他模式的一點好處是它產生的規則是比較容易懂的，舉例來說，如果薪資大於 \$30,000 和 負債比率小於 28%，則提供額外的信用額度，另外一個好處是它可以處理缺值的資料。決策樹節點支援自動和互動式訓練，當選擇自動訓練，它會自動決定輸入變數的強度，而互動式訓練則可以已經驗法則去產生決策數。

Hastie et al.提到目前被廣泛使用的決策樹演算法包括 C4.5、CHAID(Chi-Square Automatic Interaction Detector)、CART(Classification and Regression Trees)等。在使用決策樹方法進行資料分析時，其主要的優點在於能透過樹狀圖的方式，將資料的分析結果完整的呈現出來，因此在資料探勘的過程中決策樹具有視覺化與容易解釋的特色。

決策樹分類技術已經成功的被應用在不同領域，包括銀行金融、社會與行銷研究、醫療醫學等。例如陳亭羽與賀千盈(2005)利用 C4.5 決策樹應用於電子行銷市場之區隔；郭一聰(2005)利用 CART 於應收帳款之呆帳預警模式；沈偉立(2007)利用 CART 於臺股指數期貨及國民生產毛額之時間序列之預測。

### 2.2 類神經網路(Neural Network)

生物的神經網路維奇腦內的重要結構，經由神經節之間的訊息傳導，於是生物得以對許多的變化來作反應。由生物神經網路所得到的靈感，人工智慧學者嘗試著以電腦系統來模擬這樣的架構，以求增進電腦的能力，使之能夠具有生物所擁有的思考、判斷、組織等等能力。近年來由於資訊科技的進步，將類神經網路帶進另一個高峰，無論在學術研究或其他應用領域均有良好的成果表現。

類神經網路可借助本身的學習力、訓練力，由輸入的資料中歸納出隱含關係，一旦學習完成之後，以各神經元間連結強弱程度來表示，並且將此關係儲存於個人工神經元中。如果再遇到相似的輸入，就會自動依照所學習成的關係，來產生最近似的輸出，就算是資料不完整或是混亂的時候，類神經網路也會作出最佳的預測。所以類神經網路和傳統的統計方法不同，類神經網路不需要藉助是先建立好的程序和知識，它具備一般化(generalization)的能力，可以自動找出資料中所隱含的模式或特徵。

## 2.3 線性迴歸模式(Linear Regression)

迴歸分析(Regression Analysis)是一種統計分析方法，它利用一組自變數的數值，對某一應變數作預測，它也可以作為評估預測變數對準則變數的效用。也就是利用一組預測變數（或稱獨立變數或解釋變數）對某一準則變數（或稱反應變數）建立關係式，以便作為預測的依據。探討迴歸的主要目的是作預測，目標是發展一種能以一個或多個預測變數的數值來作為應變數預測的方法。例如：市場調查員想知道某產品價格與需求量的關係，則他可以藉由迴歸分析從價格來預測需求量；或者研究人員想知道新生兒身高與體重的關係，也可借用迴歸分析去探討身高與體重的關係。

將變數分為兩類，一類變數為自變數(Independent Variable)，以  $X$  表示，另一類是我們真正想知道的結果，稱為應變數(Dependent Variable)，以  $Y$  表示。所謂建立模式，就是找出  $Y$  與  $X$  的函數關係式，另外亦可利用迴歸分析來剔除預測能力不佳之自變數。

一般線性迴歸模式公式如：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{ip-1} + \varepsilon_i, i = 1, 2, \dots, n.$$

其中  $Y_i$  為應變數； $X_i$  為自變數。因為  $X$  之值需要事先知道，所以視為參數， $Y$  才是真正的變數。該母體皆為  $Y$  所組成。在此假設下，應變數之期望值與自變數之間存在一直線關係，且各個條件母體之變動狀況一致。自變數透過迴歸係數  $\beta_0, \beta_1, \dots, \beta_{p-1}$  影響應變數之期望值為邊際的影響。其他自變數是經由斜率影響應變數， $\varepsilon_i, i=1, 2, \dots, n$  為服從  $N(0, \sigma^2)$  的獨立隨機誤差項。

## 2.4 支撐向量機(Support Vector Machine)

支撐向量機(Support Vector Machine, SVM)最早起源於 60 年代，直到 1996 年 Vapnik 和 AT&T 實驗室團隊所提出之後才開始受到矚目，主要理論是來自統計學習理論終結構化風險最小誤差(Structural Risk Minimization, SRM)，目前學術研究中廣泛的應用於分類問題上。此外，運用於各個領域，文字辨識、生物資訊、網路安全、財務金融等分類問題上。

支撐向量機(SVM)是近年來被提出用於預測的方法之一，主要是根據統計學習理論為基礎，透過結構化風險最小誤差法則(SRM)原理的學習演算法。支撐向量迴歸的主要思想是針對二元分類問題，在高維度空間中尋找一個超平面作為二類的分割，以保證最小的分類錯誤率，而且 SVM 一個重要的優點就是能處理線性不可分的情況。支撐向量迴歸式將自變數與應變數間的對應關係從原本較低維度提高到高維度的特徵空間中，透過這個方法尋找一個新的對應函數，使投射出的預測效果最佳，如圖 2.4-1 所示。

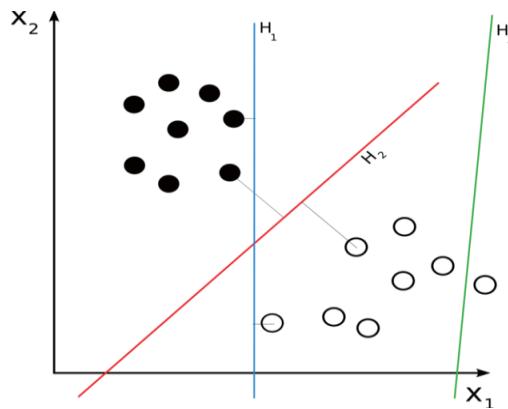


圖 2.4-1 : SVM 超平面分割示意圖

### 三、專案流程

#### 3.1 應用 SAS Enterprise Guide 整合資料及變數延伸

##### 3.1.1 原始資料

以下為 SAS 公司以及玉山銀行所提供的 5 筆資料集：

###### 3.1.1.1 顧客基本資料(SCHOOL\_CUSTOMER CST)

TABLE NAME : SCHOOL_CUSTOMER(顧客基本資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
CST_001	性別	Nominal	F/M
CST_002	子女數	Nominal	子女數量
CST_003	教育程度	Nominal	博士、碩士、大學、專科、高中(職)、國中、小學以及其他
CST_004	婚姻狀態	Nominal	已婚、未婚、離婚、再婚以及其他
CST_005	現有工作年數	Interval	職場工作時間
CST_006	收入	Interval	年收入狀況
CST_007	職業	Nominal	礦業及土石採取業、製造業、水電燃氣業、營建業、批發及零售業、住宿及餐飲業、運輸倉儲及通信業、金融及保險業、不動產及租賃業、軍/警/消、公務人員、專業科學及技術服務業、醫療保險服務業、文化運動及休閒服務業、其他服務業、學生、教育機構
CST_008	是否擁有附卡	Binary	顧客是否擁有附卡(0:無;1:有)

CST_009	擁有正卡數	Interval	正卡持有數
CST_010	擁有附卡數	Interval	附卡持有數
CST_011	分期消費註記	Binary	否持有分期消費(0:無;1:有)
CST_012	分期金註記	Binary	是否持有分期金(0:無;1:有)
CST_013	撥薪戶註記	Binary	是否持有撥薪戶(0:無;1:有)
CST_014	證券戶註記	Binary	是否持有證券戶(0:無;1:有)
CST_015	風險註記(H/M/L)	Nominal	識別顧客為何種風險客戶，分為高風險、中風險以及低風險(H/M/L)
CST_016	信用卡總額度	Interval	信用卡可消費之最大上限金額
CST_017	當月帳單金額	Interval	當月消費金額
CST_018	當月繳款金額	Interval	當月需繳款金額
CST_019	全體金融機構當月信用卡額度	Interval	除玉山銀行外之信用卡可消費之最大上限金額
CST_020	全體金融機構當月信用卡帳單金額	Interval	除玉山銀行外之信用卡可消費之最大上限金額

表 3.1-1：顧客基本資料，共 21 欄。

### 3.1.1.2 行內行為資料(SCHOOL\_CCD)

TABLE NAME : SCHOOL_CCD(行內行為資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
CCD_001	ID	Nominal	顧客識別號碼
CCD_002	日期註記	Nominal	日期識別，2011 年 2 月至 7 月，六個月間之識別註記
CCD_003	科目	Nominal	共 3 類，分別為一般消費、預借現金以及上期未繳金額(循環)
CCD_004	餘額	Interval	消費行為所剩餘額
CCD_005	次數	Interval	消費次數

表 3.1-2：行內行為資料，共 5 欄。

### 3.1.1.3 行外型為資料(SCHOOL\_ULC)

TABLE NAME : SCHOOL_ULC(行外行為資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ULC_001	ID	Nominal	顧客識別號碼
ULC_002	日期註記	Nominal	日期識別，2011 年 2 月至 7 月，六個月間之識別註記
ULC_003	科目	Nominal	共 3 類，分別為一般消費、預借現金以及上期未繳金額(循環)
ULC_004	餘額	Interval	消費行為所剩餘額

表 3.1-3：行外行為資料，共 4 欄。

### 3.1.1.4 產品持有資料(SCHOOL\_PRODUCT\_HOLDING PH)

TABLE NAME : SCHOOL_PRODUCT_HOLDING(產品持有資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
PH_001	ID	Nominal	顧客識別號碼
PH_002	產品編號	Nominal	共 8 類，分別為 1 存款、2 基金、3 其他理財、4 信貸、5 房貸、6 其他無擔保授信、7 其他擔保授信、8 其他授信。
PH_003	帳戶餘額	Interval	顧客帳戶之餘額
PH_004	帳戶狀態	Nominal	共 3 類，分別為 A 正常、B 存款靜止,授信催收呆帳,理財申購中、C 結清或贖回

表 3.1-4：產品持有資料，共 4 欄。

### 3.1.1.5 預測目標變數(SCHOOL\_TARGET)

TABLE NAME : SCHOOL_TARGET(預測目標變數)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
TARGET	預測目標變數註記-行銷活動實際成交客戶	Binary	行銷活動實際成交客戶(1:成交;0:無成交)

表 3.1-5：預測目標變數，共 2 欄。

## 3.1.2 延伸變數-行內行為資料

本研究將此表資料做合併以及延展，運用 SAS Enterprise Guide 做原始資料之合併延伸。產生兩種延伸變數資料集，如表 3.1-6 及表 3.1-7 所示。

**Type 1.** 各個 ID 共對應 18 筆資料(6 個月且分別有 3 種科目)，將一般消費科目和預借現金科目之餘額與次數分別加總合併，而上期未繳金額(循環)科目則刪取最後一個月之金額。此時，各個 ID 共對應 5 筆資料(半年且有三種科目，其中一般消費與預借現金科目又含金額與次數之狀態)。再將此 5 筆資料做為各個 ID 的欄位延伸，分別為，CCD00301(行內一般消費餘額)、CCD00301C(行內一般消費次數)、CCD00302(行內預借現金餘額)、CCD00302C(行內預借現金次數)、CCD00303(行內上期未繳金額)，共 1800000 筆資料轉變為各個 ID 含 6 個欄位之 100000 筆資料。如表 3.1-6 所示：

TABLE NAME : SCHOOL_CCD(行內行為資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
CCD00301	行內一般消費金額	Interval	2011 年 2 月至 7 月間之一般消費金額
CCD00301C	行內一般消費次數	Interval	2011 年 2 月至 7 月間之一般消費次數
CCD00302	行內預借現金金額	Interval	2011 年 2 月至 7 月間之預借現金金額
CCD00302C	行內預借現金次數	Interval	2011 年 2 月至 7 月間之預借現金次數
CCD00303	行內上期未繳金額 (循環)	Interval	截至 2011 年 7 月之未繳金額

表 3.1-6：行內行為資料 Type1

**Type 2.** 將原始資料之每月註記根據各 ID 之消費行為延伸 25 欄位(共 6 個月且分別有 2 種科目(一般消費、預借現金)與兩種狀態(金額、消費次數)，而上期未繳金額(循環)科目則如表 3.1-6，刪取最後一個月之金額。)，共 1800000 筆資料轉變為各個 ID 含 26 個欄位之 100000 筆資料。如表 3.1-7 所示：

TABLE NAME : SCHOOL_CCD(行內行為資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
CCD_2_8_A_Money	行內一般消費金額(2月)	Interval	顧客2011年2月一般消費金額
CCD_2_8_A_Count	行內一般消費次數(2月)	Interval	顧客2011年2月一般消費次數
CCD_3_9_A_Money	行內一般消費金額(3月)	Interval	顧客2011年3月一般消費金額
CCD_3_9_A_Count	行內一般消費次數(3月)	Interval	顧客2011年3月一般消費次數
CCD_4_10_A_Money	行內一般消費金額(4月)	Interval	顧客2011年4月一般消費金額
CCD_4_10_A_Count	行內一般消費次數(4月)	Interval	顧客2011年4月一般消費次數
CCD_5_11_A_Money	行內一般消費金額(5月)	Interval	顧客2011年5月一般消費金額
CCD_5_11_A_Count	行內一般消費次數(5月)	Interval	顧客2011年5月一般消費次數
CCD_6_12_A_Money	行內一般消費金額(6月)	Interval	顧客2011年6月一般消費金額
CCD_6_12_A_Count	行內一般消費次數(6月)	Interval	顧客2011年6月一般消費次數
CCD_7_1_A_Money	行內一般消費金額(7月)	Interval	顧客2011年7月一般消費金額
CCD_7_1_A_Count	行內一般消費次數(7月)	Interval	顧客2011年7月一般消費次數
CCD_2_8_B_Money	行內預借現金金額(2月)	Interval	顧客2011年2月預借現金金額
CCD_2_8_B_Count	行內預借現金次數(2月)	Interval	顧客2011年2月預借現金次數
CCD_3_9_B_Money	行內預借現金金額(3月)	Interval	顧客2011年3月預借現金金額
CCD_3_9_B_Count	行內預借現金次數(3月)	Interval	顧客2011年3月預借現金次數
CCD_4_10_B_Money	行內預借現金金額(4月)	Interval	顧客2011年4月預借現金金額
CCD_4_10_B_Count	行內預借現金次數(4月)	Interval	顧客2011年4月預借現金次數
CCD_5_11_B_Money	行內預借現金金額(5月)	Interval	顧客2011年5月預借現金金額
CCD_5_11_B_Count	行內預借現金次數(5月)	Interval	顧客2011年5月預借現金次數
CCD_6_12_B_Money	行內預借現金金額(6月)	Interval	顧客2011年6月預借現金金額
CCD_6_12_B_Count	行內預借現金次數(6月)	Interval	顧客2011年6月預借現金次數
CCD_7_1_B_Money	行內預借現金金額(7月)	Interval	顧客2011年7月預借現金金額
CCD_7_1_B_Count	行內預借現金次數(7月)	Interval	顧客2011年7月預借現金次數
CCD_7_1_C_Money	行內上期未繳金額(7月)	Interval	截至2011年7月之未繳金額

表 3.1-7：行內行為資料 Type2

### 3.1.3 延伸變數-行外行為資料

本研究將此表資料做合併以及延展，產生兩種延伸變數資料集，如表 3.1-8 及表 3.1-9 所示。

**Type 1.** 此部分之做法與行內行為資料之表 3.1-6 產出方式雷同，各個 ID 共對應 18 筆資料(6 個月且分別有 3 種科目)，將一般消費科目和預借現金科目之餘額分別加總合併，而上期未繳金額(循環)科目則刪取最後一個月之金額。此時，各個 ID 共對應 3 筆資料(半年且有三種科目)。再將此 3 筆資料做為各個 ID 的欄位延伸，分別為，ULC00301 (行外全體金融機

構一般消費金額)、ULC00302(行外全體金融機構預借現金金額)、  
ULC00303(行外全體金融機構上期未繳金額(循環))，共 1800000 筆資料  
轉變為各個 ID 含 4 個欄位之 100000 筆資料。如表 3.1-8 所示：

TABLE NAME : SCHOOL_CCD(行外行為資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
ULC00301	全體金融機構一般消費金額	Interval	2011 年 2 月至 7 月間之一般消費金額
ULC00302	全體金融機構預借現金金額	Interval	2011 年 2 月至 7 月間之預借現金金額
ULC00303	全體金融機構上期未繳金額(循環)	Interval	截至 2011 年 7 月之未繳金額

表 3.1-8：行外行為資料 Type1

**Type 2.** 將原始資料之每月註記根據各 ID 之消費行為延伸 13 欄位(共 6 個月  
且分別有 2 種科目(一般消費、預借現金)，而上期未繳金額(循環)科目則  
如表 3.1-8，刪取最後一個月之金額。)，共 1800000 筆資料轉變為各個  
ID 含 14 個欄位之 100000 筆資料。如表 3.1-9 所示：

TABLE NAME : SCHOOL\_CCD(行外行為資料)

COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
ID	ID	Nominal	顧客識別號碼
ULC_2_8_A_Money	全體金融機構一般消費金額 (2月)	Interval	顧客 2011 年 2 月一般消費金額
ULC_3_9_A_Money	全體金融機構一般消費金額 (3月)	Interval	顧客 2011 年 3 月一般消費金額
ULC_4_10_A_Money	全體金融機構一般消費金額 (4月)	Interval	顧客 2011 年 4 月一般消費金額
ULC_5_11_A_Money	全體金融機構一般消費金額 (5月)	Interval	顧客 2011 年 5 月一般消費金額
ULC_6_12_A_Money	全體金融機構一般消費金額 (6月)	Interval	顧客 2011 年 6 月一般消費金額
ULC_7_1_A_Money	全體金融機構一般消費金額 (7月)	Interval	顧客 2011 年 7 月一般消費金額
ULC_2_8_B_Money	全體金融機構預借現金金額 (2月)	Interval	顧客 2011 年 2 月預借現金金額
ULC_3_9_B_Money	全體金融機構預借現金金額 (3月)	Interval	顧客 2011 年 3 月預借現金金額
ULC_4_10_B_Money	全體金融機構預借現金金額 (4月)	Interval	顧客 2011 年 4 月預借現金金額
ULC_5_11_B_Money	全體金融機構預借現金金額 (5月)	Interval	顧客 2011 年 5 月預借現金金額
ULC_6_12_B_Money	全體金融機構預借現金金額 (6月)	Interval	顧客 2011 年 6 月預借現金金額
ULC_7_1_B_Money	全體金融機構預借現金金額 (7月)	Interval	顧客 2011 年 7 月預借現金金額
ULC_7_1_C_Money	全體金融機構上期未繳金額 (循環)	Interval	截至 2011 年 7 月之未繳金額

表 3.1-9：行外行為資料 Type2

### 3.1.4 延伸變數-產品持有資料

本研究將此表資料做合併以及延展，產生兩種延伸變數資料集，如表 3.1-10 及表 3.1-11 所示。

**Step 1.** 此表當中，相同 ID 可能有一筆產品或多筆產品，也有可能無持有產品。故此部分首先針對 8 種產品(分別為存款、基金、其他理財、信貸、房貸、其他無擔保授信、其他擔保授信、其他授信。) 以及 3 種帳戶狀態(分別為正常、存款靜止,授信催收呆帳,理財申購中、結清或贖回。) 延伸為 24 欄，各為每帳戶之不同產品之狀態與其金額，如表 3.1-10 所

示。

TABLE NAME : SCHOOL\_PRODUCT\_HOLDING(產品持有資料)

COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
PH_001	ID	Nominal	顧客識別號碼
PRODUCT_1A	存款(正常)	Interval	
PRODUCT_2A	基金(正常)	Interval	
PRODUCT_3A	其他理財(正常)	Interval	
PRODUCT_4A	信貸(正常)	Interval	
PRODUCT_5A	房貸(正常)	Interval	
PRODUCT_6A	其他無擔保(正常)授信	Interval	
PRODUCT_7A	其他擔保授信(正常)	Interval	
PRODUCT_8A	其他授信(正常)	Interval	
PRODUCT_1B	存款(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_2B	基金(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_3B	其他理財(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_4B	信貸(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_5B	房貸(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_6B	其他無擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_7B	其他擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_8B	其他授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_1C	存款(結清或贖回)	Interval	
PRODUCT_2C	基金(結清或贖回)	Interval	
PRODUCT_3C	其他理財(結清或贖回)	Interval	
PRODUCT_4C	信貸(結清或贖回)	Interval	
PRODUCT_5C	房貸(結清或贖回)	Interval	
PRODUCT_6C	其他無擔保授信(結清或贖回)	Interval	
PRODUCT_7C	其他擔保授信(結清或贖回)	Interval	
PRODUCT_8C	其他授信(結清或贖回)	Interval	

表 3.1-10：產品持有資料 Type1

Step 2. 沿用表 3.1-10，再延伸出 24 欄，表示各產品在三種狀態下是否無缺

值，無缺值表示此帳戶持有某狀態之產品，無缺值表示持有產品，以 1 表示，缺值表示未持有產品，以 0 表示，如表 3.1-11 所示。

TABLE NAME : SCHOOL_PRODUCT_HOLDING(產品持有資料)			
COLUMN NAME	CHINESE DESC	LEVEL	DESCRIPTION
PH_001	ID	Nominal	顧客識別號碼
PRODUCT_1A	存款(正常)	Interval	
PRODUCT_2A	基金(正常)	Interval	
PRODUCT_3A	其他理財(正常)	Interval	
PRODUCT_4A	信貸(正常)	Interval	
PRODUCT_5A	房貸(正常)	Interval	
PRODUCT_6A	其他無擔保(正常)授信	Interval	
PRODUCT_7A	其他擔保授信(正常)	Interval	
PRODUCT_8A	其他授信(正常)	Interval	
PRODUCT_1B	存款(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_2B	基金(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_3B	其他理財(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_4B	信貸(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_5B	房貸(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_6B	其他無擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_7B	其他擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_8B	其他授信(存款靜止，授信催收呆帳，理財申購中)	Interval	
PRODUCT_1C	存款(結清或贖回)	Interval	
PRODUCT_2C	基金(結清或贖回)	Interval	
PRODUCT_3C	其他理財(結清或贖回)	Interval	
PRODUCT_4C	信貸(結清或贖回)	Interval	
PRODUCT_5C	房貸(結清或贖回)	Interval	
PRODUCT_6C	其他無擔保授信(結清或贖回)	Interval	
PRODUCT_7C	其他擔保授信(結清或贖回)	Interval	
PRODUCT_8C	其他授信(結清或贖回)	Interval	
PRODUCT_1Ab	存款-擁有且狀態正常(1:有;0:無)	Binary	

PRODUCT_2Ab	基金-擁有且狀態正常(1:有;0:無)	Binary	
PRODUCT_3Ab	其他理財-擁有且正常(1:有;0:無)	Binary	
PRODUCT_4Ab	信貸-擁有且正常(1:有;0:無)	Binary	
PRODUCT_5Ab	房貸-擁有且正常(1:有;0:無)	Binary	
PRODUCT_6Ab	其他無擔保授信-擁有且正常(1:有;0:無)	Binary	
PRODUCT_7Ab	其他擔保授信-擁有且正常(1:有;0:無)	Binary	
PRODUCT_8Ab	其他授信-擁有且正常(1:有;0:無)	Binary	
PRODUCT_1Bb	存款-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_2Bb	基金-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_3Bb	其他理財-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_4Bb	信貸-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_5Bb	房貸-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_6Bb	其他無擔保授信-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_7Bb	其他擔保授信-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_8Bb	其他授信-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary	
PRODUCT_1Cb	存款-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_2Cb	基金-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_3Cb	其他理財-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_4Cb	信貸-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_5Cb	房貸-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_6Cb	其他無擔保授信-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_7Cb	其他擔保授信-擁有且結清或贖回(1:有;0:無)	Binary	
PRODUCT_8Cb	其他授信-擁有且結清或贖回(1:有;0:無)	Binary	

表 3.1-11：產品持有資料 Type2

## 3.2 應用 SAS Enterprise Miner 建立預測模型

首先先針對本專案所需的資料來源及資料的選取方式做說明；接著再說明資料的前處理步驟；說明資料分割；說明利用決策樹、回歸、類神經網路及 SVM；說明訓練模型後，利用模型比較的節點來看 ROC 圖；說明評分資料的使用並且將平分資料匯入 SCORE 的節點去做預測；最後再利用 SAS CODE 將資料以 SAS 的格式匯出

### 3.2.1 輸入資料

經由資料處理後，本團隊整理出了 3 筆資料集，分別為 SAS\_0731、SAS\_0801 以及 SAS\_0803，而他們的變數個數分別為 54、78、108，如表 3.2-1、表 3.2-2 和表 3.2-3 所示。

TABLE NAME: SAS_0731			
#	COLUMN NAME	CHINESE DESC	LEVEL
1	ID	ID	Nominal
2	CST_001	性別	Nominal
3	CST_002	子女數	Nominal
4	CST_003	教育程度	Nominal
5	CST_004	婚姻狀態	Nominal
6	CST_005	現有工作年數	Interval
7	CST_006	收入	Interval
8	CST_007	職業	Nominal
9	CST_008	是否擁有附卡	Binary
10	CST_009	擁有正卡數	Interval
11	CST_010	擁有附卡數	Interval
12	CST_011	分期消費註記	Binary
13	CST_012	分期金註記	Binary
14	CST_013	撥薪戶註記	Binary
15	CST_014	證券戶註記	Binary
16	CST_015	風險註記(H/M/L)	Nominal
17	CST_016	信用卡總額度	Interval
18	CST_017	當月帳單金額	Interval
19	CST_018	當月繳款金額	Interval
20	CST_019	全體金融機構當月信用卡額度	Interval
21	CST_020	全體金融機構當月信用卡帳單金額	Interval
22	CCD00301	行內一般消費金額	Interval
23	CCD00301C	行內一般消費次數	Interval
24	CCD00302	行內預借現金金額	Interval

25	CCD00302C	行內預借現金次數	Interval
26	CCD00303	行內上期未繳金額(循環)	Interval
27	ULC00301	行外全體金融機構一般消費金額	Interval
28	ULC00302	行外全體金融機構預借現金金額	Interval
29	ULC00303	行外全體金融機構上期未繳金額(循環)	Interval
30	PRODUCT_1A	存款(正常)	Interval
31	PRODUCT_2A	基金(正常)	Interval
32	PRODUCT_3A	其他理財(正常)	Interval
33	PRODUCT_4A	信貸(正常)	Interval
34	PRODUCT_5A	房貸(正常)	Interval
35	PRODUCT_6A	其他無擔保(正常)授信	Interval
36	PRODUCT_7A	其他擔保授信(正常)	Interval
37	PRODUCT_8A	其他授信(正常)	Interval
38	PRODUCT_1B	存款(存款靜止，授信催收呆帳，理財申購中)	Interval
39	PRODUCT_2B	基金(存款靜止，授信催收呆帳，理財申購中)	Interval
40	PRODUCT_3B	其他理財(存款靜止，授信催收呆帳，理財申購中)	Interval
41	PRODUCT_4B	信貸(存款靜止，授信催收呆帳，理財申購中)	Interval
42	PRODUCT_5B	房貸(存款靜止，授信催收呆帳，理財申購中)	Interval
43	PRODUCT_6B	其他無擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval
44	PRODUCT_7B	其他擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval
45	PRODUCT_8B	其他授信(存款靜止，授信催收呆帳，理財申購中)	Interval
46	PRODUCT_1C	存款(結清或贖回)	Interval
47	PRODUCT_2C	基金(結清或贖回)	Interval
48	PRODUCT_3C	其他理財(結清或贖回)	Interval
49	PRODUCT_4C	信貸(結清或贖回)	Interval
50	PRODUCT_5C	房貸(結清或贖回)	Interval
51	PRODUCT_6C	其他無擔保授信(結清或贖回)	Interval
52	PRODUCT_7C	其他擔保授信(結清或贖回)	Interval
53	PRODUCT_8C	其他授信(結清或贖回)	Interval
54	TARGET	目標變數	Binary

表 3.2-1 : SAS\_0731 資料集

TABLE NAME: SAS_0801			
#	COLUMN NAME	CHINESE DESC	LEVEL
1	ID	ID	Nominal
2	CST_001	性別	Nominal
3	CST_002	子女數	Nominal
4	CST_003	教育程度	Nominal
5	CST_004	婚姻狀態	Nominal
6	CST_005	現有工作年數	Interval
7	CST_006	收入	Interval
8	CST_007	職業	Nominal
9	CST_008	是否擁有附卡	Binary
10	CST_009	擁有正卡數	Interval
11	CST_010	擁有附卡數	Interval
12	CST_011	分期消費註記	Binary
13	CST_012	分期金註記	Binary
14	CST_013	撥薪戶註記	Binary
15	CST_014	證券戶註記	Binary
16	CST_015	風險註記(H/M/L)	Nominal
17	CST_016	信用卡總額度	Interval
18	CST_017	當月帳單金額	Interval
19	CST_018	當月繳款金額	Interval
20	CST_019	全體金融機構當月信用卡額度	Interval
21	CST_020	全體金融機構當月信用卡帳單金額	Interval
22	CCD00301	行內一般消費金額	Interval
23	CCD00301C	行內一般消費次數	Interval
24	CCD00302	行內預借現金金額	Interval
25	CCD00302C	行內預借現金次數	Interval
26	CCD00303	行內上期未繳金額(循環)	Interval
27	ULC00301	行外全體金融機構一般消費金額	Interval
28	ULC00302	行外全體金融機構預借現金金額	Interval
29	ULC00303	行外全體金融機構上期未繳金額(循環)	Interval
30	PRODUCT_1A	存款(正常)	Interval
31	PRODUCT_2A	基金(正常)	Interval
32	PRODUCT_3A	其他理財(正常)	Interval
33	PRODUCT_4A	信貸(正常)	Interval
34	PRODUCT_5A	房貸(正常)	Interval
35	PRODUCT_6A	其他無擔保(正常)授信	Interval
36	PRODUCT_7A	其他擔保授信(正常)	Interval
37	PRODUCT_8A	其他授信(正常)	Interval

38	PRODUCT_1B	存款(存款靜止，授信催收呆帳，理財申購中)	Interval
39	PRODUCT_2B	基金(存款靜止，授信催收呆帳，理財申購中)	Interval
40	PRODUCT_3B	其他理財(存款靜止，授信催收呆帳，理財申購中)	Interval
41	PRODUCT_4B	信貸(存款靜止，授信催收呆帳，理財申購中)	Interval
42	PRODUCT_5B	房貸(存款靜止，授信催收呆帳，理財申購中)	Interval
43	PRODUCT_6B	其他無擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval
44	PRODUCT_7B	其他擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval
45	PRODUCT_8B	其他授信(存款靜止，授信催收呆帳，理財申購中)	Interval
46	PRODUCT_1C	存款(結清或贖回)	Interval
47	PRODUCT_2C	基金(結清或贖回)	Interval
48	PRODUCT_3C	其他理財(結清或贖回)	Interval
49	PRODUCT_4C	信貸(結清或贖回)	Interval
50	PRODUCT_5C	房貸(結清或贖回)	Interval
51	PRODUCT_6C	其他無擔保授信(結清或贖回)	Interval
52	PRODUCT_7C	其他擔保授信(結清或贖回)	Interval
53	PRODUCT_8C	其他授信(結清或贖回)	Interval
54	PRODUCT_1Ab	存款-擁有且狀態正常(1:有;0:無)	Binary
55	PRODUCT_2Ab	基金-擁有且狀態正常(1:有;0:無)	Binary
56	PRODUCT_3Ab	其他理財-擁有且正常(1:有;0:無)	Binary
57	PRODUCT_4Ab	信貸-擁有且正常(1:有;0:無)	Binary
58	PRODUCT_5Ab	房貸-擁有且正常(1:有;0:無)	Binary
59	PRODUCT_6Ab	其他無擔保授信-擁有且正常(1:有;0:無)	Binary
60	PRODUCT_7Ab	其他擔保授信-擁有且正常(1:有;0:無)	Binary
61	PRODUCT_8Ab	其他授信-擁有且正常(1:有;0:無)	Binary
62	PRODUCT_1Bb	存款-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
63	PRODUCT_2Bb	基金-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
64	PRODUCT_3Bb	其他理財-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
65	PRODUCT_4Bb	信貸-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
66	PRODUCT_5Bb	房貸-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
67	PRODUCT_6Bb	其他無擔保授信-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary
68	PRODUCT_7Bb	其他擔保授信-擁有且存款靜止，授信催收呆帳，理財申購中(1:有;0:無)	Binary

69	PRODUCT_8Bb	其他授信-擁有且存款靜止，授信催收呆帳，理財申購中 (1:有;0:無)	Binary
70	PRODUCT_1Cb	存款-擁有且結清或贖回(1:有;0:無)	Binary
71	PRODUCT_2Cb	基金-擁有且結清或贖回(1:有;0:無)	Binary
72	PRODUCT_3Cb	其他理財-擁有且結清或贖回(1:有;0:無)	Binary
73	PRODUCT_4Cb	信貸-擁有且結清或贖回(1:有;0:無)	Binary
74	PRODUCT_5Cb	房貸-擁有且結清或贖回(1:有;0:無)	Binary
75	PRODUCT_6Cb	其他無擔保授信-擁有且結清或贖回(1:有;0:無)	Binary
76	PRODUCT_7Cb	其他擔保授信-擁有且結清或贖回(1:有;0:無)	Binary
77	PRODUCT_8Cb	其他授信-擁有且結清或贖回(1:有;0:無)	Binary
78	TARGET	目標變數	Binary

表 3.2-2 : SAS\_0801 資料集

TABLE NAME: SAS_0803			
#	COLUMN NAME	CHINESE DESC	LEVEL
1	ID	ID	Nominal
2	CST_001	性別	Nominal
3	CST_002	子女數	Nominal
4	CST_003	教育程度	Nominal
5	CST_004	婚姻狀態	Nominal
6	CST_005	現有工作年數	Interval
7	CST_006	收入	Interval
8	CST_007	職業	Nominal
9	CST_008	是否擁有附卡	Binary
10	CST_009	擁有正卡數	Interval
11	CST_010	擁有附卡數	Interval
12	CST_011	分期消費註記	Binary
13	CST_012	分期金註記	Binary
14	CST_013	撥薪戶註記	Binary
15	CST_014	證券戶註記	Binary
16	CST_015	風險註記(H/M/L)	Nominal
17	CST_016	信用卡總額度	Interval
18	CST_017	當月帳單金額	Interval
19	CST_018	當月繳款金額	Interval
20	CST_019	全體金融機構當月信用卡額度	Interval
21	CST_020	全體金融機構當月信用卡帳單金額	Interval
22	PRODUCT_1A	存款(正常)	Interval
23	PRODUCT_2A	基金(正常)	Interval
24	PRODUCT_3A	其他理財(正常)	Interval
25	PRODUCT_4A	信貸(正常)	Interval
26	PRODUCT_5A	房貸(正常)	Interval
27	PRODUCT_6A	其他無擔保(正常)授信	Interval
28	PRODUCT_7A	其他擔保授信(正常)	Interval
29	PRODUCT_8A	其他授信(正常)	Interval
30	PRODUCT_1B	存款(存款靜止，授信催收呆帳，理財申購中)	Interval
31	PRODUCT_2B	基金(存款靜止，授信催收呆帳，理財申購中)	Interval
32	PRODUCT_3B	其他理財(存款靜止，授信催收呆帳，理財申購中)	Interval
33	PRODUCT_4B	信貸(存款靜止，授信催收呆帳，理財申購中)	Interval
34	PRODUCT_5B	房貸(存款靜止，授信催收呆帳，理財申購中)	Interval
35	PRODUCT_6B	其他無擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval
36	PRODUCT_7B	其他擔保授信(存款靜止，授信催收呆帳，理財申購中)	Interval

		中)	
37	PRODUCT_8B	其他授信(存款靜止，授信催收呆帳，理財申購中)	Interval
38	PRODUCT_1C	存款(結清或贖回)	Interval
39	PRODUCT_2C	基金(結清或贖回)	Interval
40	PRODUCT_3C	其他理財(結清或贖回)	Interval
41	PRODUCT_4C	信貸(結清或贖回)	Interval
42	PRODUCT_5C	房貸(結清或贖回)	Interval
43	PRODUCT_6C	其他無擔保授信(結清或贖回)	Interval
44	PRODUCT_7C	其他擔保授信(結清或贖回)	Interval
45	PRODUCT_8C	其他授信(結清或贖回)	Interval
46	PRODUCT_1Ab	存款-擁有且狀態正常(1:有;0:無)	Binary
47	PRODUCT_2Ab	基金-擁有且狀態正常(1:有;0:無)	Binary
48	PRODUCT_3Ab	其他理財-擁有且正常(1:有;0:無)	Binary
49	PRODUCT_4Ab	信貸-擁有且正常(1:有;0:無)	Binary
50	PRODUCT_5Ab	房貸-擁有且正常(1:有;0:無)	Binary
51	PRODUCT_6Ab	其他無擔保授信-擁有且正常(1:有;0:無)	Binary
52	PRODUCT_7Ab	其他擔保授信-擁有且正常(1:有;0:無)	Binary
53	PRODUCT_8Ab	其他授信-擁有且正常(1:有;0:無)	Binary
54	PRODUCT_1Bb	存款-擁有且存款靜止，授信催收呆帳，理財申購中 (1:有;0:無)	Binary
55	PRODUCT_2Bb	基金-擁有且存款靜止，授信催收呆帳，理財申購中 (1:有;0:無)	Binary
56	PRODUCT_3Bb	其他理財-擁有且存款靜止，授信催收呆帳，理財申 購中(1:有;0:無)	Binary
57	PRODUCT_4Bb	信貸-擁有且存款靜止，授信催收呆帳，理財申購中 (1:有;0:無)	Binary
58	PRODUCT_5Bb	房貸-擁有且存款靜止，授信催收呆帳，理財申購中 (1:有;0:無)	Binary
59	PRODUCT_6Bb	其他無擔保授信-擁有且存款靜止，授信催收呆帳， 理財申購中(1:有;0:無)	Binary
60	PRODUCT_7Bb	其他擔保授信-擁有且存款靜止，授信催收呆帳，理 財申購中(1:有;0:無)	Binary
61	PRODUCT_8Bb	其他授信-擁有且存款靜止，授信催收呆帳，理財申 購中(1:有;0:無)	Binary
62	PRODUCT_1Cb	存款-擁有且結清或贖回(1:有;0:無)	Binary
63	PRODUCT_2Cb	基金-擁有且結清或贖回(1:有;0:無)	Binary
64	PRODUCT_3Cb	其他理財-擁有且結清或贖回(1:有;0:無)	Binary
65	PRODUCT_4Cb	信貸-擁有且結清或贖回(1:有;0:無)	Binary

66	PRODUCT_5Cb	房貸-擁有且結清或贖回(1:有;0:無)	Binary
67	PRODUCT_6Cb	其他無擔保授信-擁有且結清或贖回(1:有;0:無)	Binary
68	PRODUCT_7Cb	其他擔保授信-擁有且結清或贖回(1:有;0:無)	Binary
69	PRODUCT_8Cb	其他授信-擁有且結清或贖回(1:有;0:無)	Binary
70	CCD_2_8_A_Money	行內一般消費金額(2月)	Interval
71	CCD_2_8_A_Count	行內一般消費次數(2月)	Interval
72	CCD_3_9_A_Money	行內一般消費金額(3月)	Interval
73	CCD_3_9_A_Count	行內一般消費次數(3月)	Interval
74	CCD_4_10_A_Money	行內一般消費金額(4月)	Interval
75	CCD_4_10_A_Count	行內一般消費次數(4月)	Interval
76	CCD_5_11_A_Money	行內一般消費金額(5月)	Interval
77	CCD_5_11_A_Count	行內一般消費次數(5月)	Interval
78	CCD_6_12_A_Money	行內一般消費金額(6月)	Interval
79	CCD_6_12_A_Count	行內一般消費次數(6月)	Interval
80	CCD_7_1_A_Money	行內一般消費金額(7月)	Interval
81	CCD_7_1_A_Count	行內一般消費次數(7月)	Interval
82	CCD_2_8_B_Money	行內預借現金金額(2月)	Interval
83	CCD_2_8_B_Count	行內預借現金次數(2月)	Interval
84	CCD_3_9_B_Money	行內預借現金金額(3月)	Interval
85	CCD_3_9_B_Count	行內預借現金次數(3月)	Interval
86	CCD_4_10_B_Money	行內預借現金金額(4月)	Interval
87	CCD_4_10_B_Count	行內預借現金次數(4月)	Interval
88	CCD_5_11_B_Money	行內預借現金金額(5月)	Interval
89	CCD_5_11_B_Count	行內預借現金次數(5月)	Interval
90	CCD_6_12_B_Money	行內預借現金金額(6月)	Interval
91	CCD_6_12_B_Count	行內預借現金次數(6月)	Interval
92	CCD_7_1_B_Money	行內預借現金金額(7月)	Interval
93	CCD_7_1_B_Count	行內預借現金次數(7月)	Interval
94	CCD_7_1_C_Money	杭內預借現金金額(7月)	Interval
95	ULC_2_8_A_Money	行外全體金融機構一般消費金額(2月)	Interval
96	ULC_3_9_A_Money	行外全體金融機構一般消費金額(3月)	Interval
97	ULC_4_10_A_Money	行外全體金融機構一般消費金額(4月)	Interval
98	ULC_5_11_A_Money	行外全體金融機構一般消費金額(5月)	Interval
99	ULC_6_12_A_Money	行外全體金融機構一般消費金額(6月)	Interval
100	ULC_7_1_A_Money	行外全體金融機構一般消費金額(7月)	Interval
101	ULC_2_8_B_Money	行外全體金融機構預借現金金額(2月)	Interval
102	ULC_3_9_B_Money	行外全體金融機構預借現金金額(3月)	Interval
103	ULC_4_10_B_Money	行外全體金融機構預借現金金額(4月)	Interval

104	ULC_5_11_B_Money	行外全體金融機構預借現金金額(5月)	Interval
105	ULC_6_12_B_Money	行外全體金融機構預借現金金額(6月)	Interval
106	ULC_7_1_B_Money	行外全體金融機構預借現金金額(7月)	Interval
107	ULC_7_1_C_Money	行外全體金融機構上期未繳金額循環之金額(2月)	Interval
108	TARGET	目標變數	Binary

表 3.2-3 : SAS\_0803 資料集

### 3.2.2 資料切割

為建立一個預測模型，須將資料切割成訓練、驗證及測試的三個樣本，訓練樣本是做為模型建立所需的資料；驗證樣本則是在模式建立過程裡用來修正預測模型，使其更精確，測試樣本則是在模型建立完成後，用來評估模型的正確性，此部分本小組將資料切割成訓練及驗證兩部分。

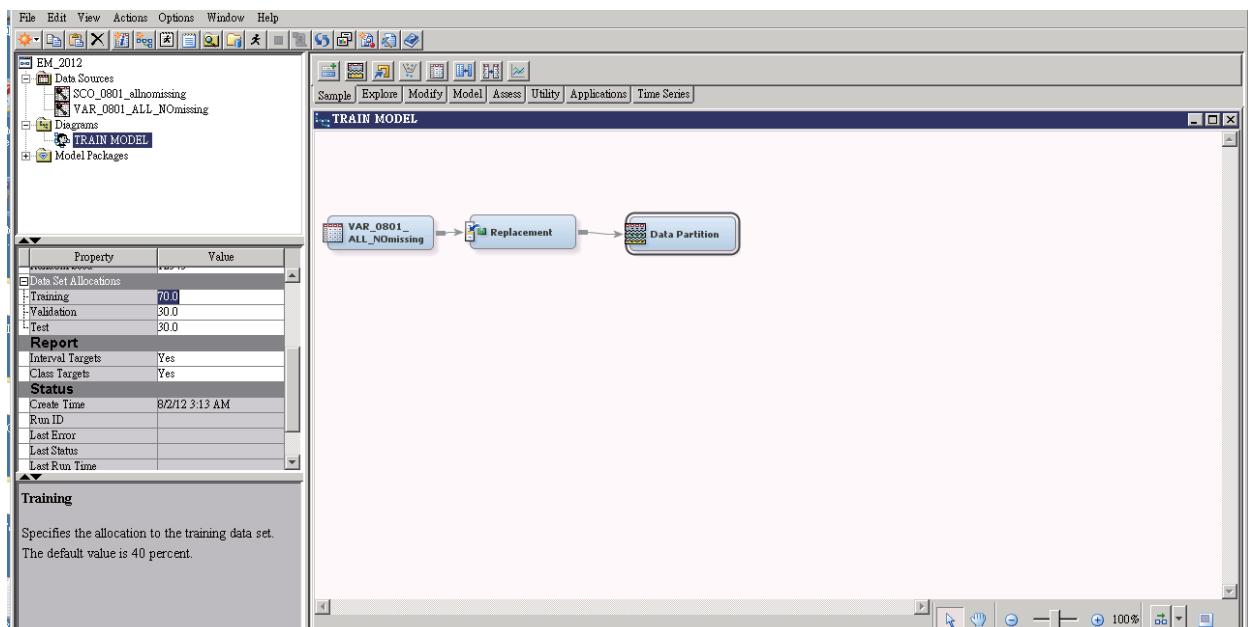


圖 3.2-1：資料切割

在資料分割中，將資料集分割成訓練及驗證兩部分，根據過往的研究中，發現將比例調乘 7:3 是最好的訓練方式，故資料集比分別例調整為 70% 和 30%，如圖 3.2-2 所示。

Property	Value
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	

Property	Value
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	

圖 3.2-2 : Data Partition 參數調整

### 3.2.3 模型建立

#### 3.2.3.1 決策樹

在建立決策樹的過程中，本小組使用兩個決策樹來建立模型，一個為系統自動建立的樹，另一個為本小組手動建立的樹。自建樹部分，本小組使用預設值執行，以下為此部分之步驟以及說明：

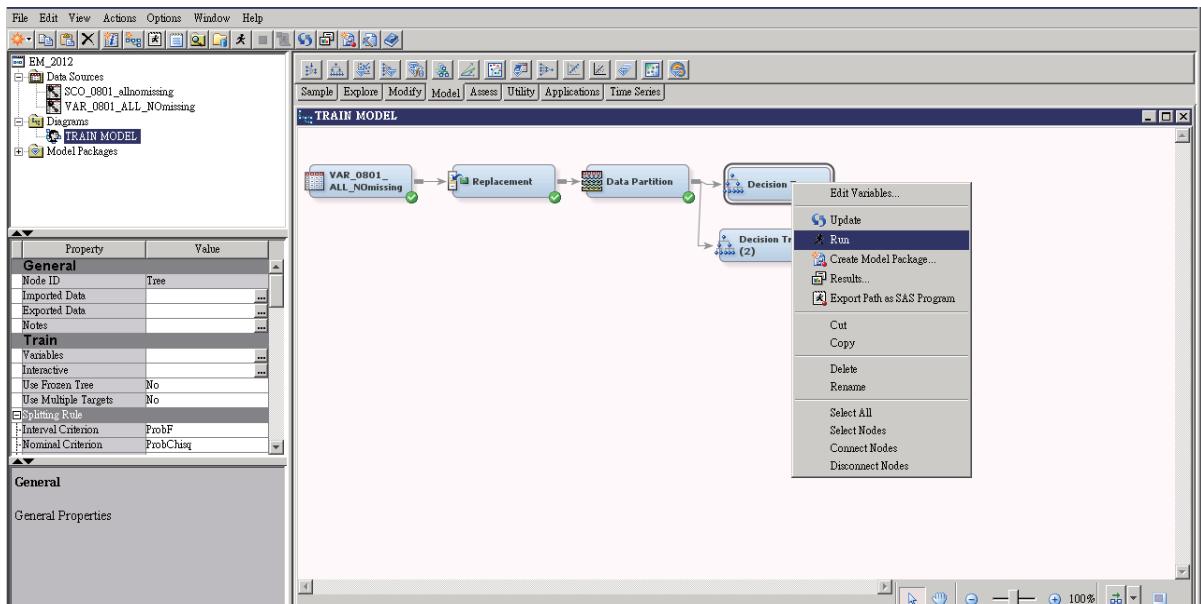


圖 3.2-3 : 自建樹執行

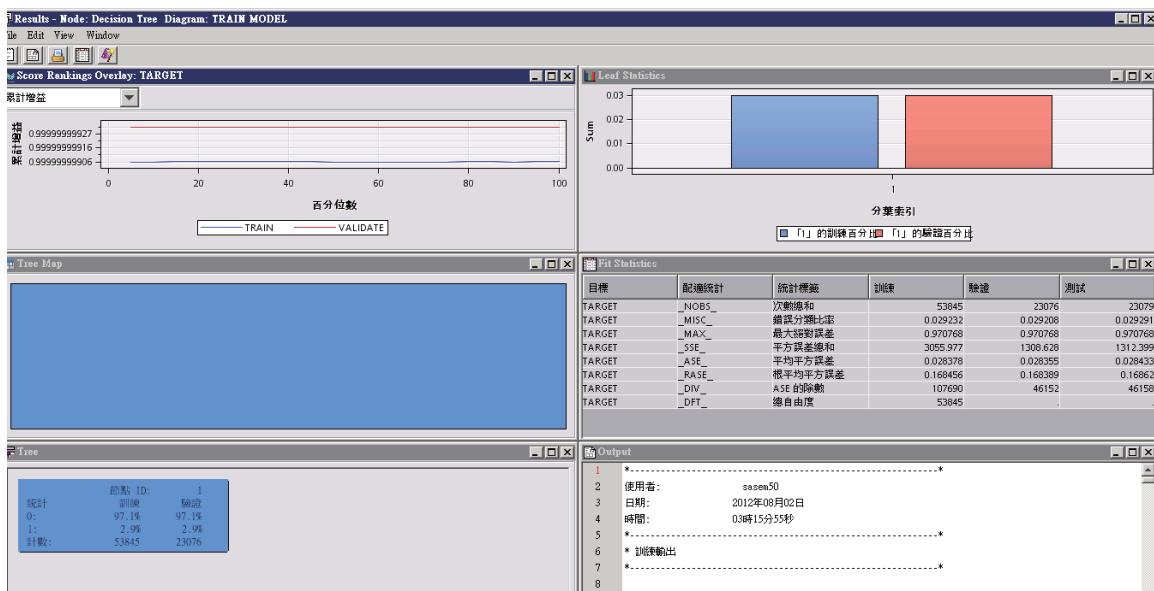


圖 3.2-4：自建樹輸出結果

圖 3.2-4 為輸出之結果，結果發現，此決策樹並無做分割的動作，表示自建樹建議不需要分割。

手動樹部分，本小組針對 $-\log(p)$ 的大小去做分割，由於 p-value 越小越好，而所產出的值太小，顧以取 LOG 將 p-value 做轉換，如此一來，數值越小就會變成越大，但因數值小於 1 時，LOG 會呈現負數，為了將解決此問題，最終就取 $-\log(p)$ 做呈現。

$-\log(p)$ 越大越具有影響力，在本研究中，先針對資料筆數多的跟 $-\log(p)$ 大的做分割依據，直到節點的 $-\log(p)$ 小於 5 時，就不再往下做分割。本小組是根據過往文獻研究及小組內部討論後決議取 $-\log(5) < 5$ 就停止分割。以下為此部分之步驟以及說明：

選擇 INTERACTIVE 建立自建樹，如圖 3.2-5 所示。分割之步驟依據 $-\log(p)$ ，先選擇資料量多的，再選擇 $-\log(p)$ 大的，當 $-\log(p)$ 小於 5 就停止，如圖 3.2-6 所示。

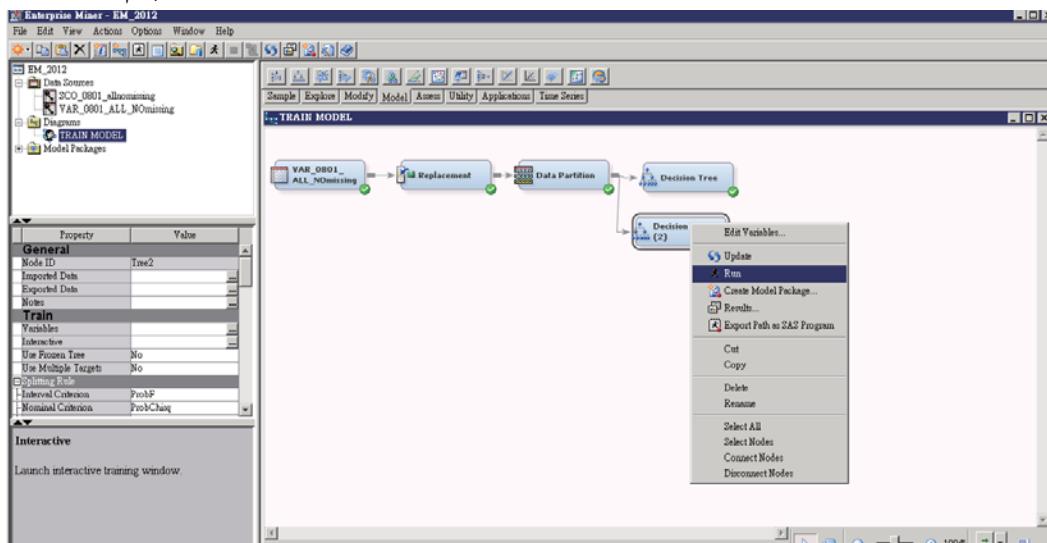


圖 3.2-5：手動樹執行

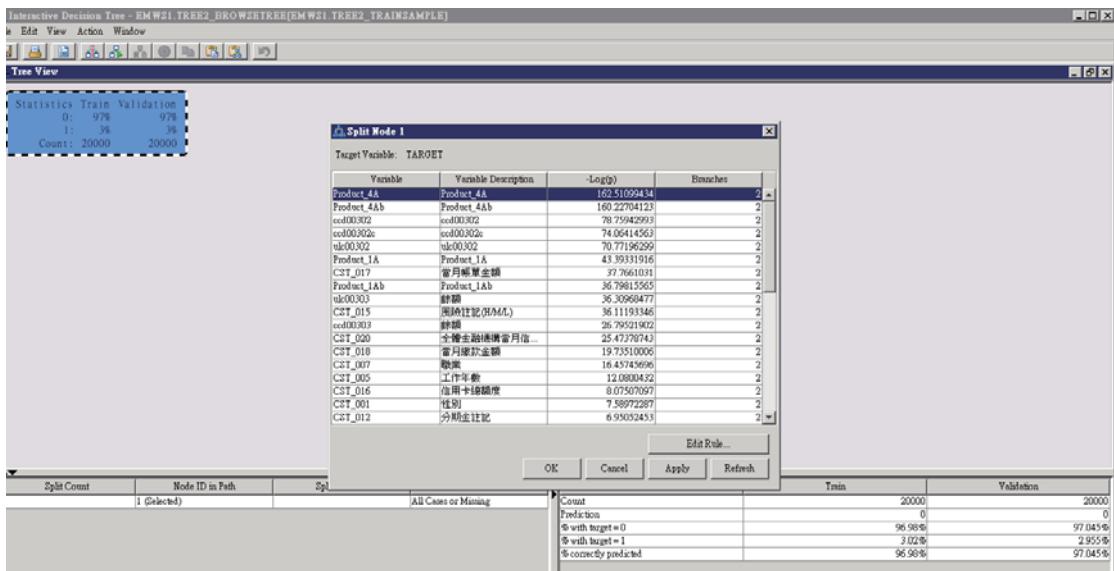


圖 3.2-6：手動樹節點分割

### 3.2.3.2 迴歸

建立迴歸之前，須對資料做前置處理，因使用迴歸時，若資料集中有遺失值，對模型可能造成很大的影響，針對此部分，本小組使用 SAS-EM 中的 IMPUTE 節點將資料集中的遺失值以其他值取代，根據過往的文獻研究中，發現平均值是最常用以及最好的取代值，而此部分本小組使用平均值做取代，如圖 3.2-7 與圖 3.2-8 所示。

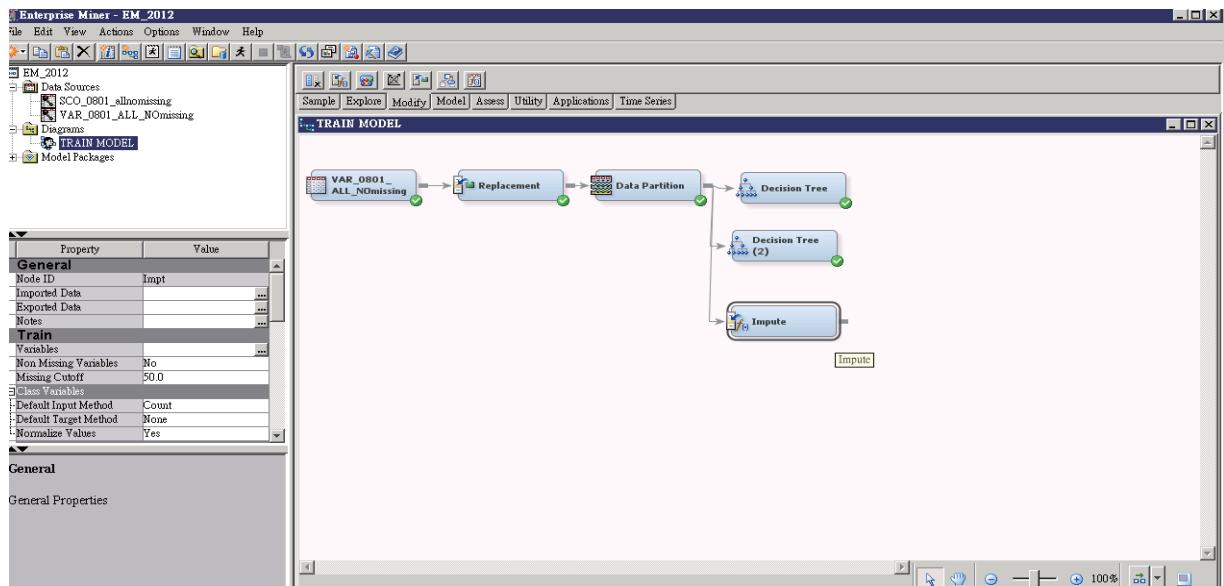


圖 3.2-7：插入 Impute 節點

Property	Value
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
<b>Score</b>	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
<b>Report</b>	
Validation and Test Data	No
Distribution of missing	No

Property	Value
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
<b>Score</b>	
Hide Original Variables	Yes
Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input
<b>Report</b>	
Validation and Test Data	No
Distribution of missing	No

圖 3.2-8 : Impute 參數調整

經過本小組討論後，使用 stepwise 方法，因綜合 forward 與 backward，且經過反覆測試後，發現其預測能力最佳，故本組經過多次討論後選擇 stepwise，如圖 3.2-9 與圖 3.2-10 所示。調整後，點選執行，圖 3.2-11 為迴歸之輸出結果。

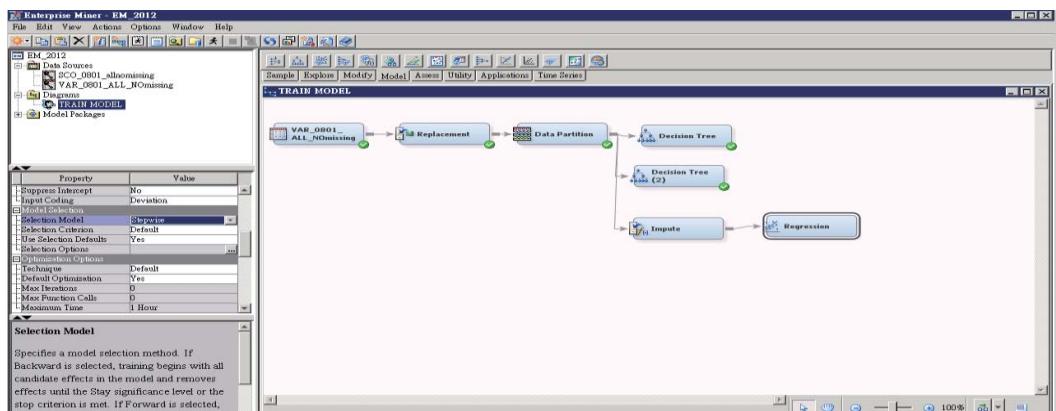


圖 3.2-9: 迴歸參數使用逐步回歸

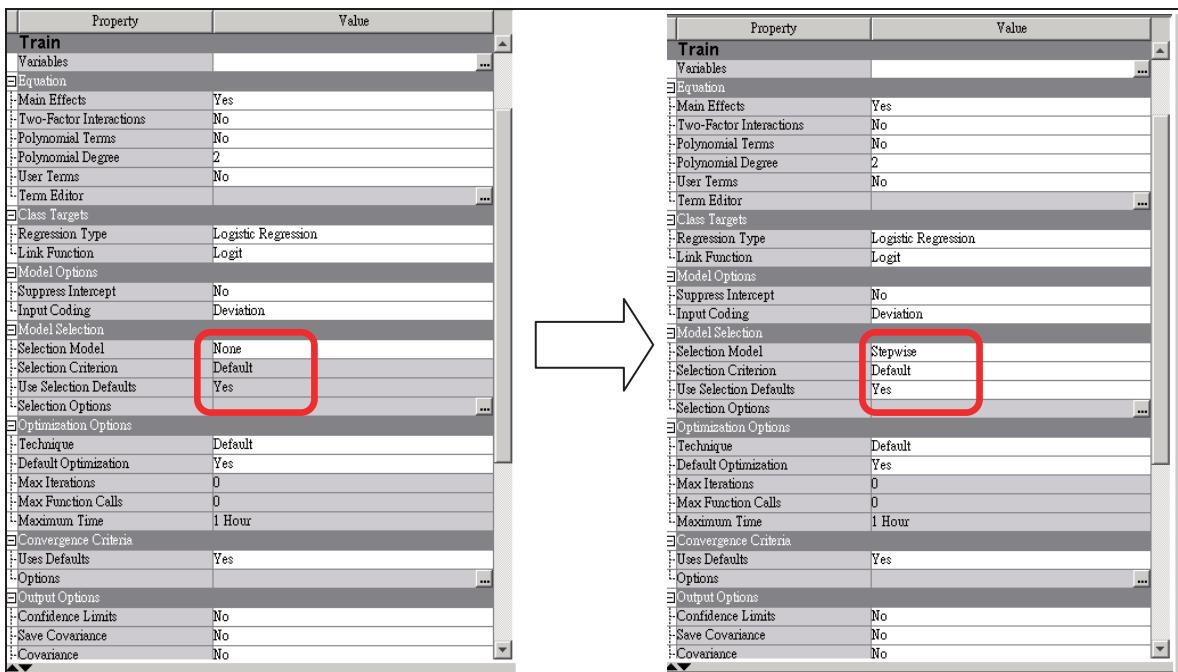


圖 3.2-10: 迴歸參數調整

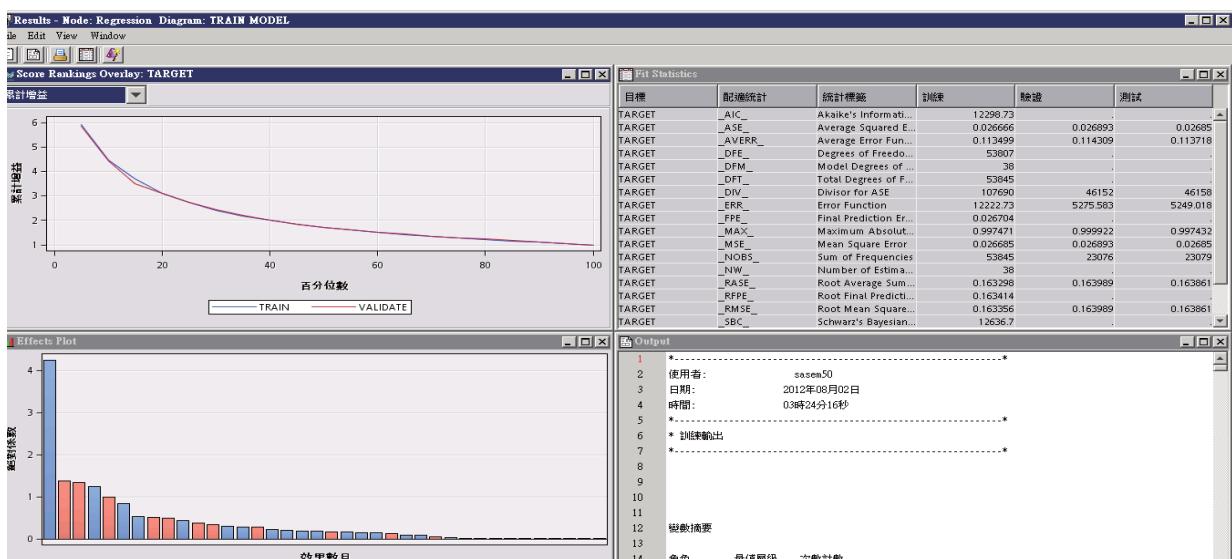


圖 3.2-11: 迴歸輸出結果。

### 3.2.3.3 類神經

類神經分三層級，分別為輸入層、隱藏層以及輸出層。類神經為迴歸之衍伸，從  $x$  跟  $y$  兩個變數延伸至  $n$  個變數，故此部分將類神經節點接於迴歸節點後做模型的執行，如圖 3.2-12 所示：

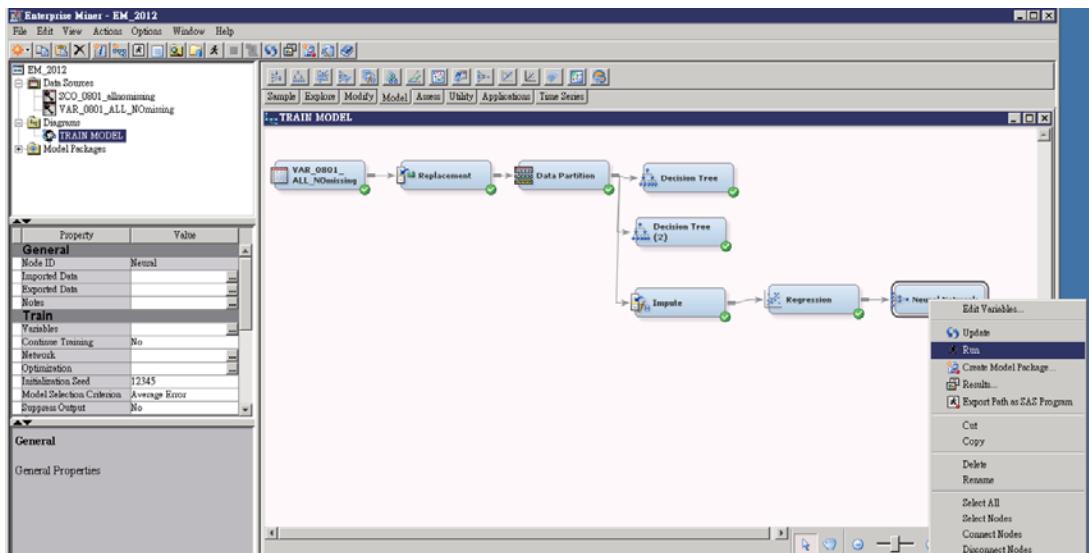


圖 3.2-12：插入類神經節點

類神經模型之選擇的準則中，本小組根據過去的研究以及反覆測試發現，運用平均誤差方法的效果最佳，故將此部分之利潤損失改為平均誤差，如圖 3.2-13 所示，圖 3.2-14 為類神經執行後之輸出結果。

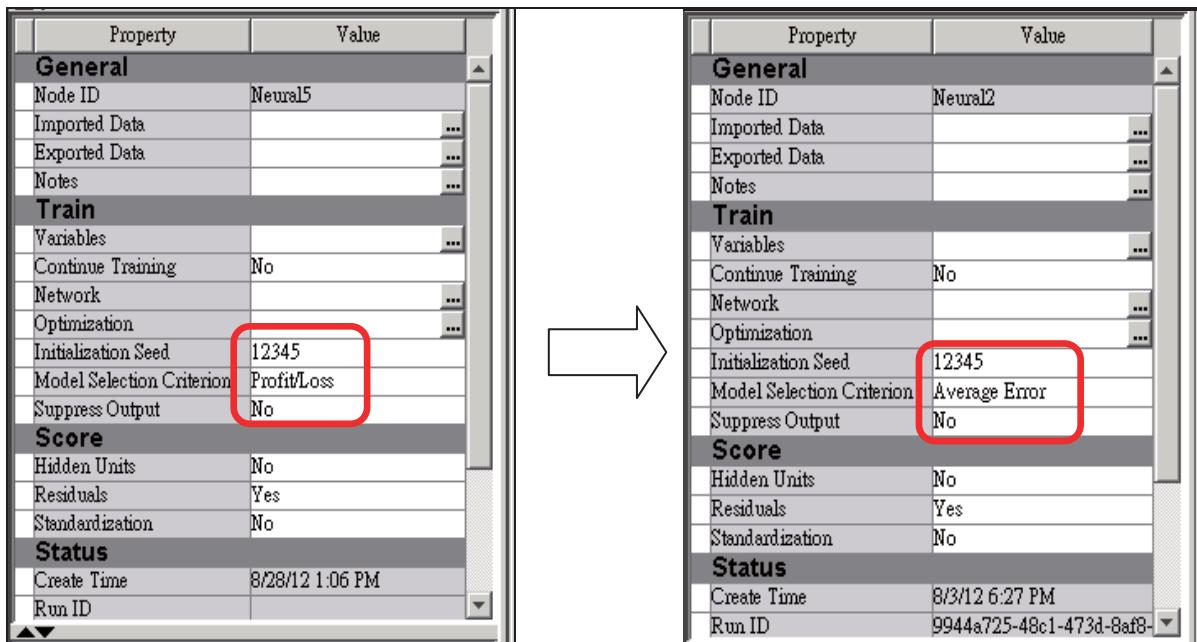


圖 3.2-13: 類神經參數調整

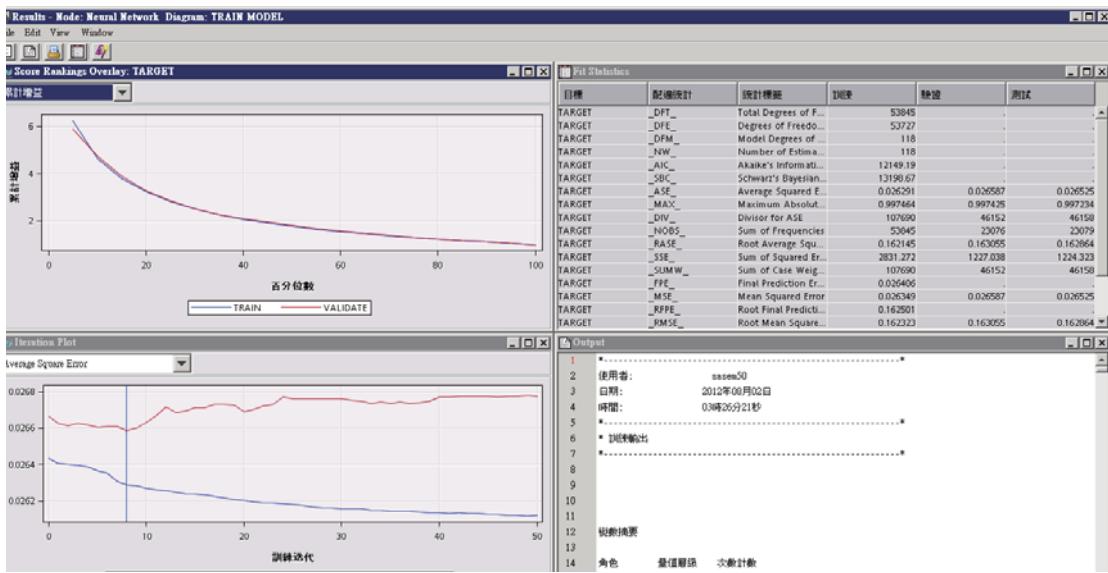


圖 3.2-14: 類神經輸出結果

### 3.2.3.4 SVM

本研究中，本小組嘗試了 SVM 的模型(如圖 3.2-15)，當中的參數，使用的估計方法為 LSVM，核心算法是使用線性。參數選擇中，本小組測試了所有參數的組合，經反覆測試後，本小組發現 LSVM 的線性方式為最佳結果，故此 SVM 模型使用的估計方法為 LSVM，其核心算法則使用線性，如圖 3.2-16 所示，調整後執行，輸出結果如圖 3.2-17。

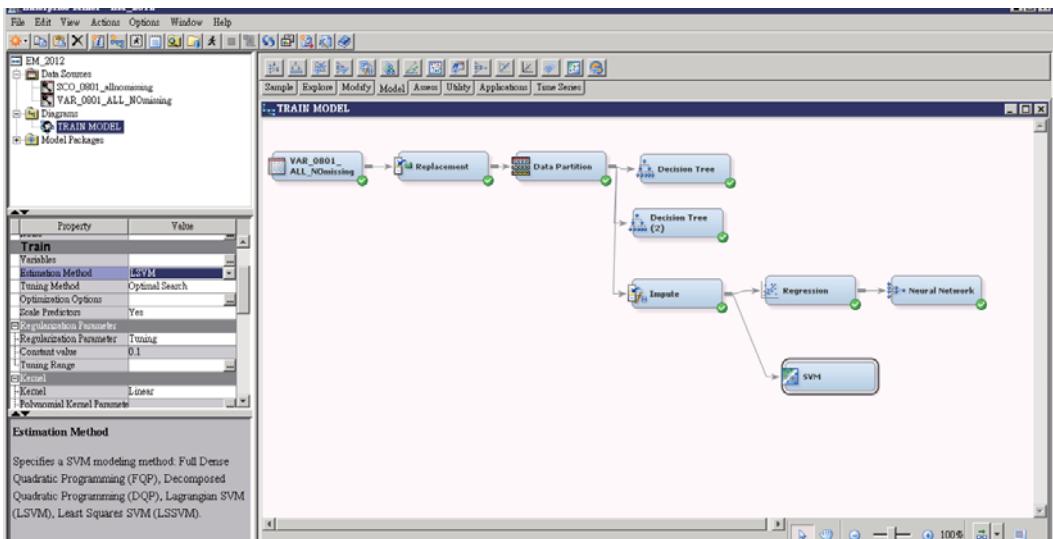


圖 3.2-15：插入 SVM 節點

The figure shows two side-by-side SVM parameter configuration tables. A red box highlights the 'Estimation Method' field in both tables, which is changed from 'DQP' in the left table to 'LSVM' in the right table. An arrow points from the left table to the right table.

Property	Value
<b>Train</b>	
Variables	
Estimation Method	DQP
Tuning Method	Optimal Search
Optimization Options	
Scale Predictors	Yes
Regularization Parameter	
Regularization Parameter	Tuning
Constant value	0.1
Tuning Range	
Kernel	
Kernel	Linear
Polynomial Kernel Parameter	
RBF Kernel Parameter	
Sigmoid Kernel Parameter	
Cross Validation	
Cross Validation	No
Method	Random
Fold	Default
Sampling	
Apply Sampling	Yes
Sample Size	1000
Print Options	
Print Option	Default
Optimization History	No

Property	Value
<b>Train</b>	
Variables	
Estimation Method	LSVM
Tuning Method	None
Optimization Options	
Scale Predictors	Yes
Regularization Parameter	
Regularization Parameter	Constant
Constant value	0.1
Tuning Range	
Kernel	
Kernel	Linear
Polynomial Kernel Parameter	
RBF Kernel Parameter	
Sigmoid Kernel Parameter	
Cross Validation	
Cross Validation	Yes
Method	Random
Fold	Default
Sampling	
Apply Sampling	Yes
Sample Size	1000
Print Options	
Print Option	Default
Optimization History	No

圖 3.2-16: SVM 參數調整

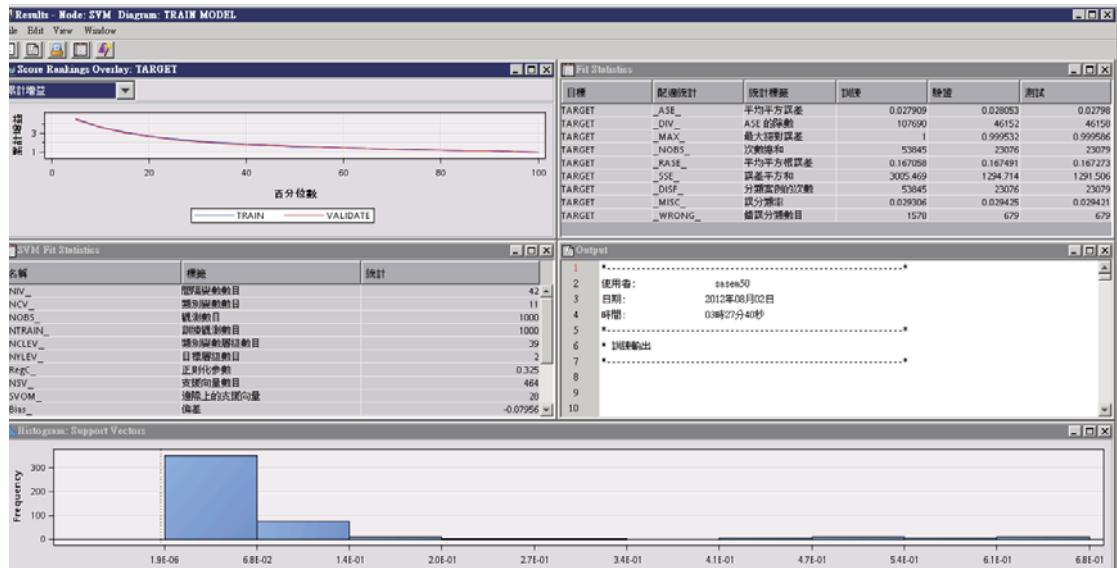


圖 3.2-17 : SVM 輸出結果

### 3.2.4 模型比較

將所有模型執行訓練後，運用模型比較節點來比較本研究中所使用的 5 個模型的好壞，如圖 3.2-18 所示，根據執行後之結果，利用 ROC 圖來做觀察，比較模型好壞，根據圖 3.2-19 顯示類神經之預測結果較佳。另外還可觀察累計增益表，了解每個模型之 LIFT 值提升多少，如圖 3.2-20。

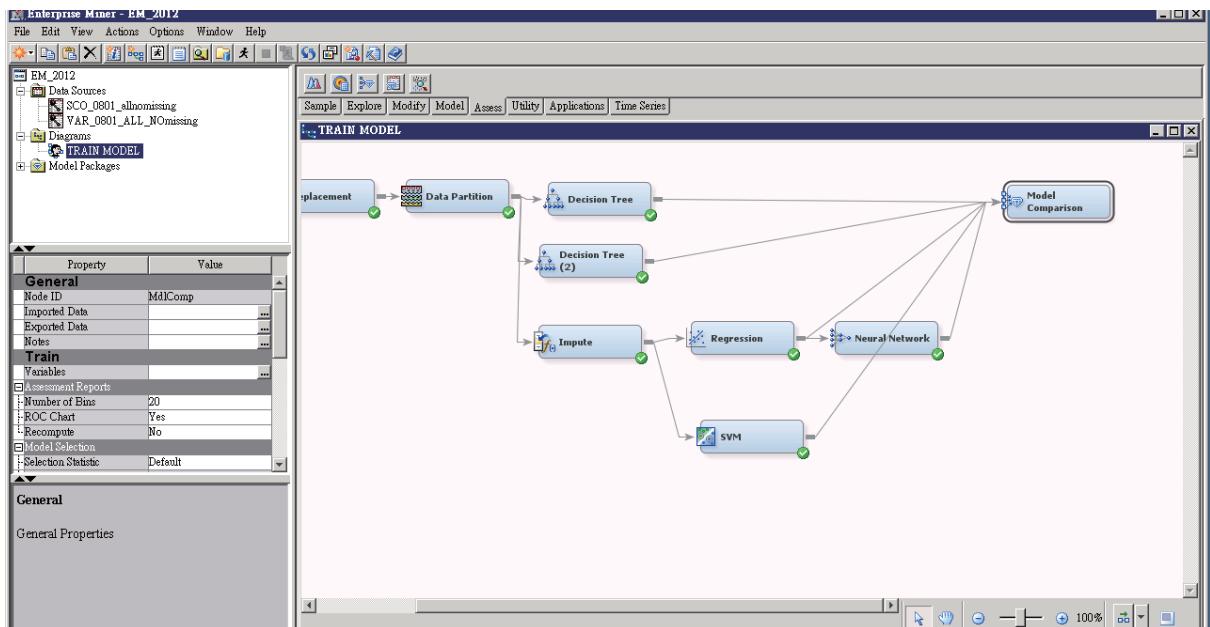


圖 3.2-18: 模型比較

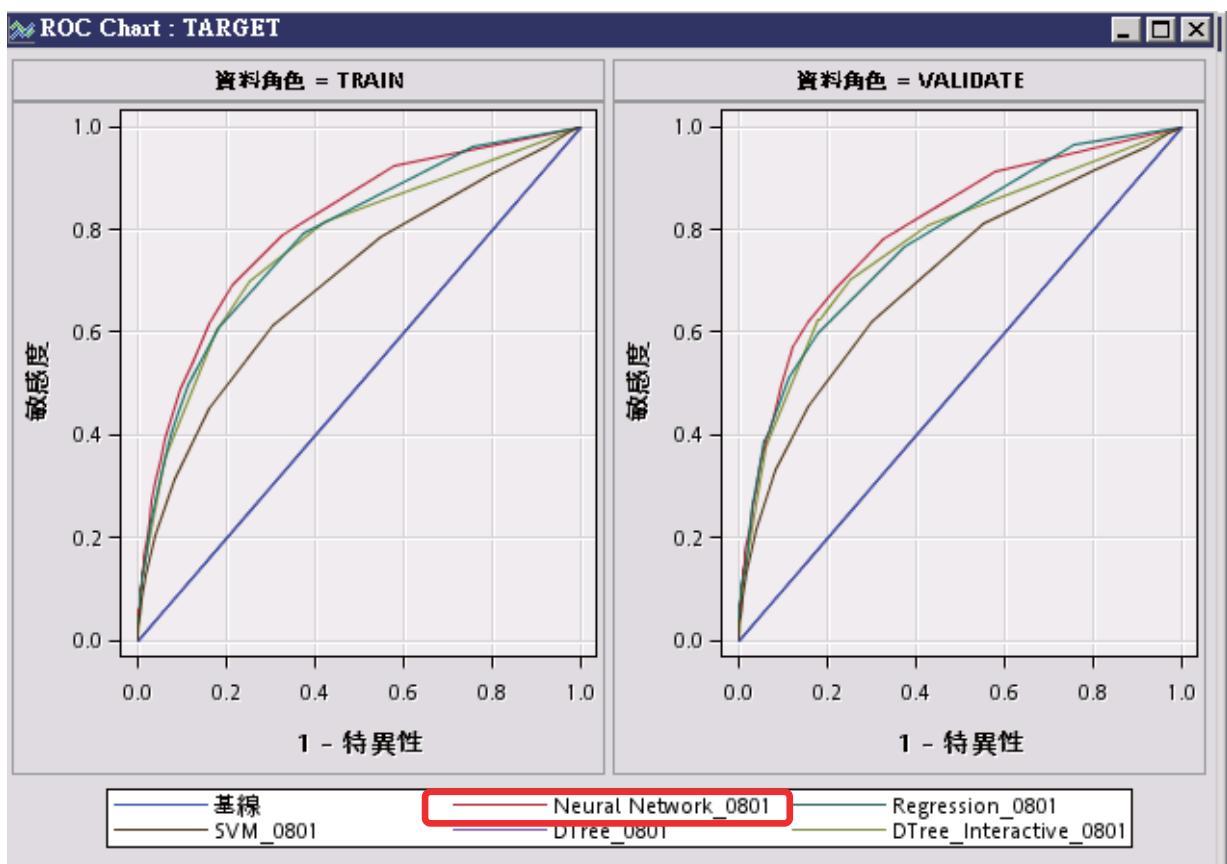


圖 3.2-19: ROC 圖

模型描述	訓練: 累計 增益	驗證: 累計 增益
Neural Network_0801	4.638681	4.65766
Regression_0801	4.37941	4.513338
SVM_0801	3.314148	3.489965
DTree_0801	1	1
DTree_Interactive_0801	4.098597	4.324223

圖 3.2-20: 累計增益表

### 3.2.5 評分資料

所有模型執行並訓練完後，將 SAS 公司及玉山銀行所提供的評分資料合併為一個資料集後，將此資料集匯入至 SAS-EM 中，其中資料集需選擇為評分資料，並運用 SCORE 節點，做可能為成交顧客之預測，如圖 3.2-21 所示，執行後之結果如圖 3.2-22。

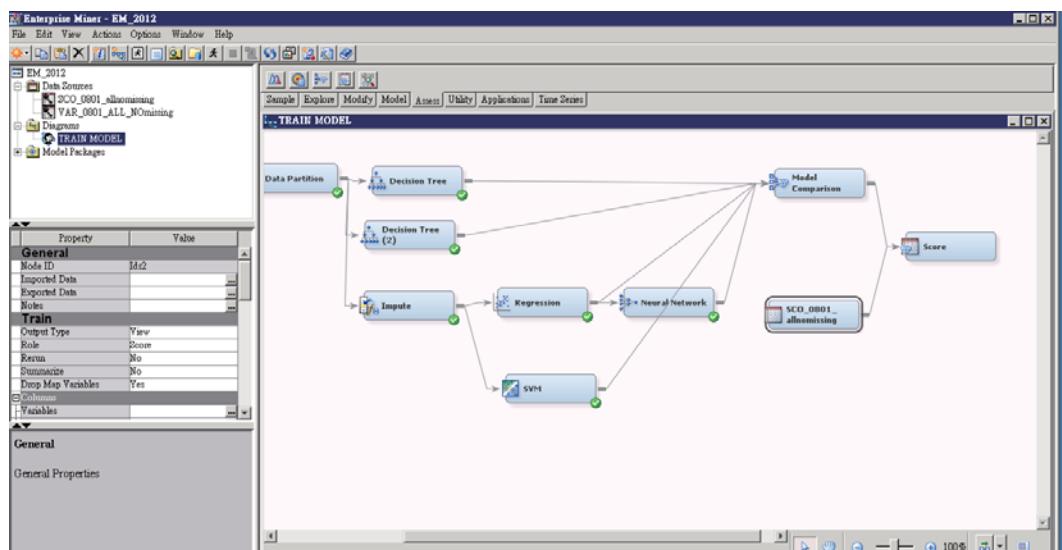


圖 3.2-21: 模型比較後方，插入 SCORE 節點。

```

' Results - Node: Score Diagram: TRAIN MODEL
de Edit View Window
File Edit View Window
Optimized SAS Code
1 * EM SCORE CODE;
2 * EM Version: 7.1;
3 * SAS Release: 9.03.OIMIP110211;
4 * Host: NO1337610168574;
5 * Encoding: ms-950;
6 * Locale: zh_TW;
7 * Project Path: E:\sasem50\result;
8 * Project Name: EM_2012;
9 * Diagram Id: EM01;
10 * Diagram Name: TRAIN MODEL;
11 * Generated by: sasem50;
12 * Date: 02/08/2012:03:31:19;
13 * ...
14 * ...
15 * TOOL: Input Data Source;
16

SAS Code
1 * EM SCORE CODE;
2 * EM Version: 7.1;
3 * SAS Release: 9.03.OIMIP110211;
4 * Host: NO1337610168574;
5 * Encoding: ms-950;
6 * Locale: zh_TW;
7 * Project Path: E:\sasem50\result;
8 * Project Name: EM_2012;
9 * Diagram Id: EM01;
10 * Diagram Name: TRAIN MODEL;
11 * Generated by: sasem50;
12 * Date: 02/08/2012:03:31:19;
13 * ...
14 * ...
15 * ...
16

Output
1 * 使用者: sasem50
2 * 日期: 2012/08/02日
3 * 時間: 03時31分33秒
4 * ...
5 * ...
6 * 調整輸出
7 * ...
8 *
9 *
10 *
11 *
12 * 变数摘要
13 *
14 * 角色 量值属性 次數計數
15 *

Output Variables
變數名稱 建立者 變數標籤 函數 線型
EM_CLASSIFICATION Score Prediction for TARGET CLASSIFICATION C
EM_EVENTPROBABILITY Score Probability for level 1 ... PREDICT N
EM_PROBABILITY Score Probability of Classific... PREDICT N
EM_SEGMENT Score 區段 TRANSFORM N
IMP_CST_005 Impt Imputed: 工作年資 TRANSFORM N
IMP_CST_015 Impt Imputed: 風險註記IV... TRANSFORM C
IMP_CST_016 Impt Imputed: 風險註記V... TRANSFORM N
IMP_CST_018 Impt Imputed: 風險註記VI... TRANSFORM N
IMP_CST_019 Impt Imputed: 風險註記VII... TRANSFORM N
IMP_UC00301 Impt Imputed uc00301 TRANSFORM N
IMP_UC00302 Impt Imputed uc00302 TRANSFORM N
IMP_UC00303 Impt Imputed uc00303 TRANSFORM N
F_TARGET Reg Info: TARGET Classification C
P_TARGET0 Predicted: TARGET=0 PREDICT N
P_TARGET1 Predicted: TARGET=1 PREDICT N

```

圖 3.2-22: SCORE 執行結果

檢視 SCORE 節點執行後之資料，其中將會包含依據 MODEL 所推論出之預測之結果資訊，檢視步驟如圖 3.2-23 與圖 3.2-24 所示，

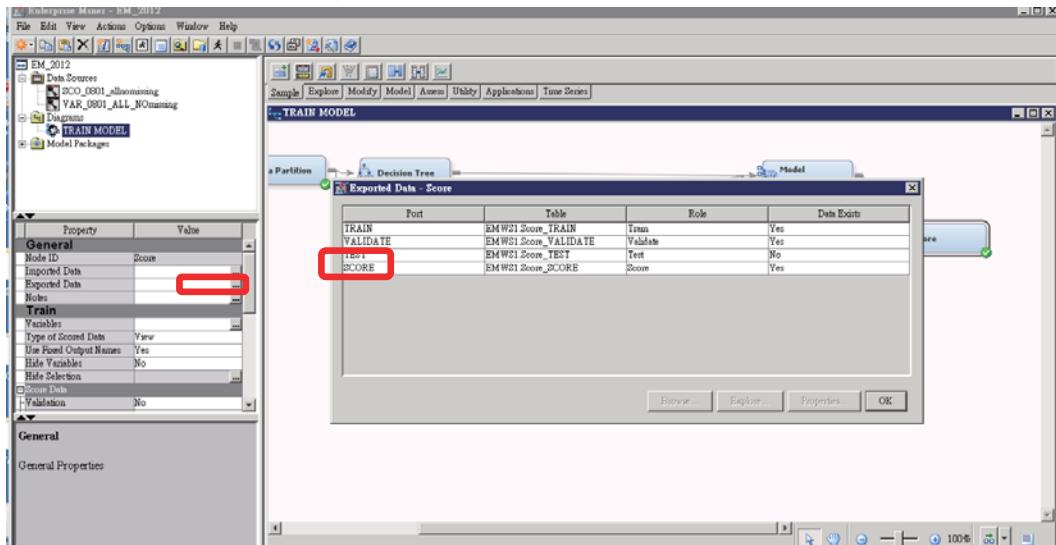


圖 3.2-23：檢視 SCORE 節點執行後結果資訊-點選 EXPORY DATA

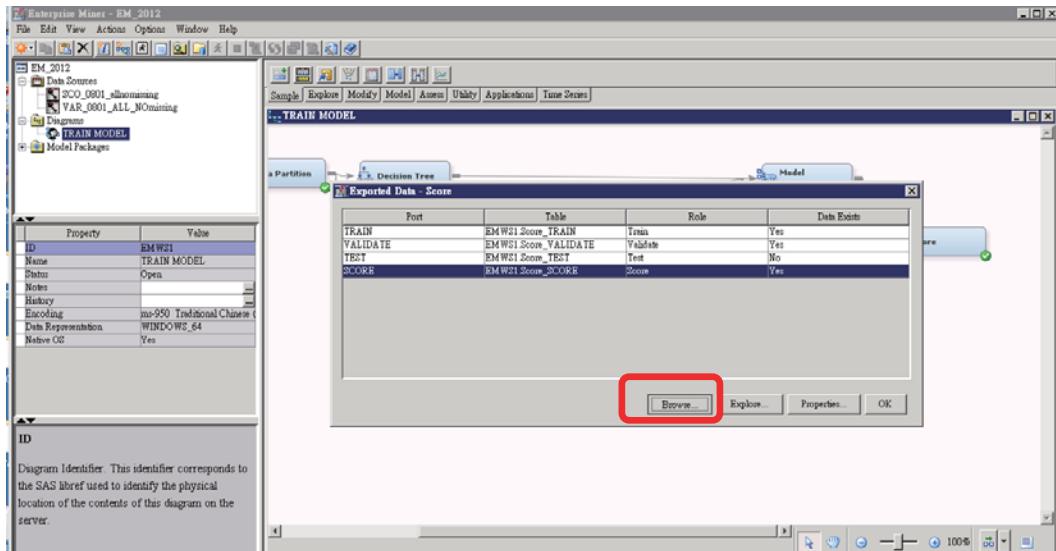


圖 3.2-24：檢視 SCORE 節點執行後結果資訊-點選 Browse。

此部分可以運用 EXPORY DATA 此功能，觀察所有 ID 經過 MODEL 推論後，預測 TARGET 為 1 的機率，點選後，將能打開有輸出結果之視窗，如圖 3.2-25 所示，將視窗拖曳至後段欄位，將能觀察所有 ID 預測 TARGET 為 1 的機率，如圖 3.2-26 所示。

ID	性別	子女數	教育程度	婚姻狀態	工作年數	年收入	職業	是否轉介財付	轉介正卡數	轉介財卡數	本行分期金註記	本行繳帳金註記	本行透支戶註記
1	0	0	8	2	0.0	428560.0	C06	0	2.0	0.0	0	0	0
2	0	4	1	0.0	835992.0	C15	0	2.0	0.0	0	0	0	
3	0	2	2	2.0	400492.0	C04	1	3.0	2.0	1	0	0	
4	0	4	1	0.0	409996.0	C10	0	1.0	0.0	0	0	0	
5	0	3	2	1.0	293328.0	C12	0	6.0	0.0	0	0	0	
6	0	5	2	1.66	180000.0	C05	0	2.0	0.0	0	0	0	
7	0	3	1	0.0	951996.0	C12	1	9.0	2.0	0	0	0	
8	0	3	2	0.0	319992.0	C15	0	1.0	0.0	0	0	0	
9	0	5	1	0.0	399996.0	C12	0	1.0	0.0	0	0	0	
10	0	3	2	0.33	0.0	C02	0	2.0	0.0	0	0	0	
11	0	4	1	0.0	100000.0	C06	0	1.0	0.0	0	0	0	
12	0	3	2	1.41	564996.0	C02	0	1.0	0.0	0	0	1	
13	0	3	2	0.0	322692.0	C15	0	1.0	0.0	0	0	0	
14	0	1	2	2.5	400492.0	C17	0	1.0	0.0	0	0	0	
15	0	1	2	2.5	400492.0	C17	0	1.0	0.0	0	0	0	
16	0	3	2	0.66	331200.0	C06	0	1.0	0.0	0	0	1	
17	0	5	1	4.0	159990.0	C10	0	2.0	0.0	0	0	0	
18	0	4	2	0.0	259992.0	C14	0	1.0	0.0	1	0	1	
19	0	5	2	1.0	259992.0	C02	0	2.0	0.0	0	0	0	
20	0	3	2	0.0	199992.0	C17	0	2.0	0.0	1	0	1	
21	0	5	2	9.0	1165800.0	C10	0	1.0	0.0	0	0	1	
22	0	5	2	6.0	180000.0	C15	0	1.0	0.0	0	0	0	
23	0	2	2	0.16	159990.0	C05	0	2.0	0.0	0	0	1	
24	0	3	2	1.25	199992.0	C12	0	2.0	0.0	0	0	1	
25	0	4	2	0.0	986798.0	C02	0	1.0	0.0	0	0	1	
26	0	5	1	2.5	120000.0	C02	0	1.0	0.0	1	0	0	
27	0	3	2	4.0	399996.0	C10	0	1.0	0.0	0	0	0	
28	0	8	1	0.0	322692.0	C15	0	1.0	0.0	0	0	1	
29	0	3	2	0.0	345996.0	C15	0	3.0	0.0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	

圖 3.2-25: SCORE 輸出結果資訊

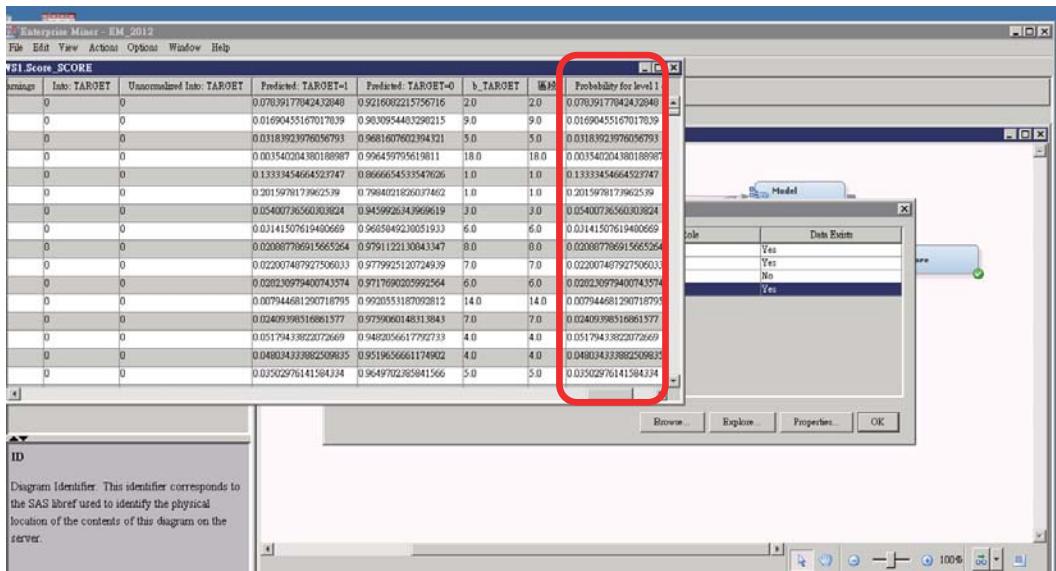


圖 3.2-26: 觀察 TARGET 為 1 之資訊

### 3.2.6 匯出

運用 SAS CODE 節點做預測後之資料匯出，如圖 3.2-27 表示：

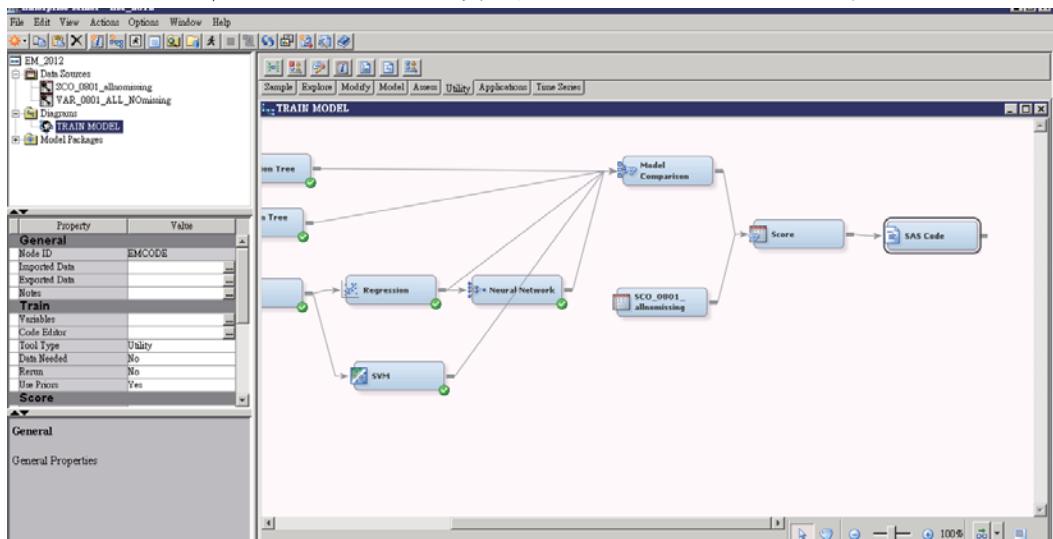


圖 3.2-27: 插入 SAS CODE 節點

在此節點之屬性欄位中，點選 CODE EDITOR 來做匯出程式的撰寫。依照程式之格式，撰寫完畢後於左上方點選儲存，即可執行。此時評分資料將會匯至 LIBRARY 中，如圖 3.2-28 所示。

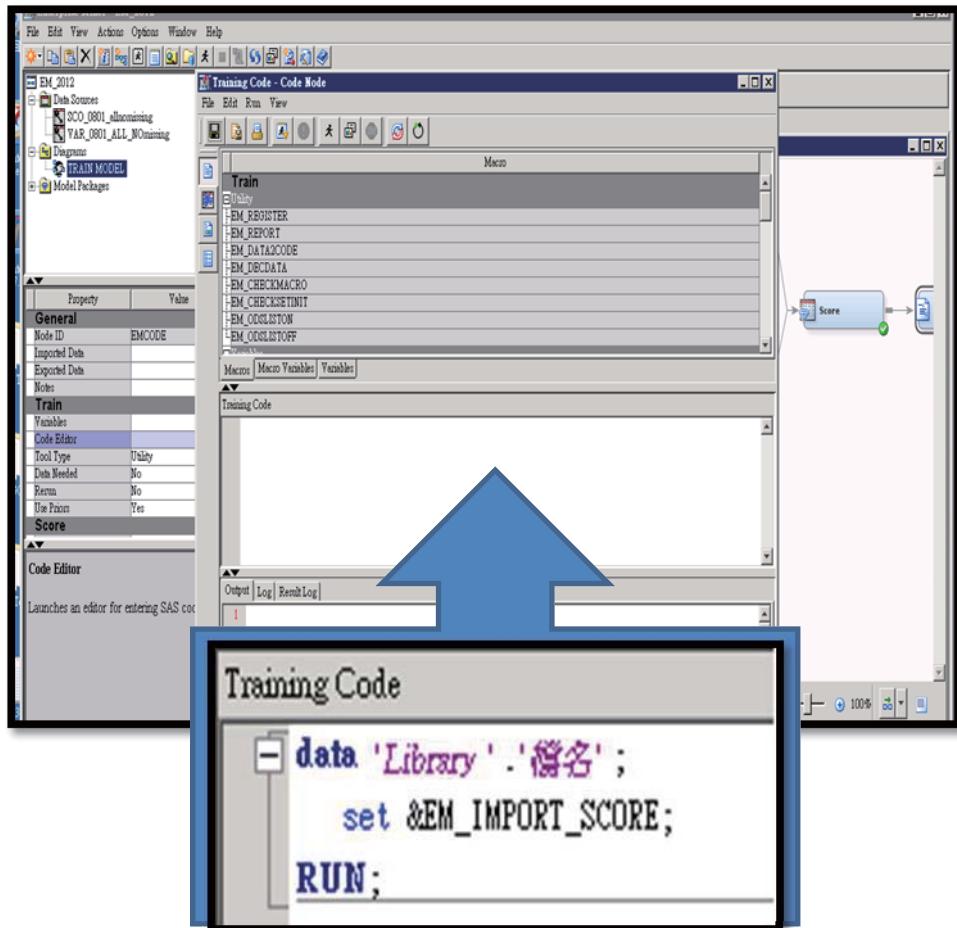


圖 3.2-28:匯出 CODE 撰寫

存放目錄中即可找到 SAS-EM 評分後所匯出之 5 萬筆評分資料，將此資料至 SAS-EG 中做 TARGET=1 機率遞減排序並取前 10000 名之銷售名單。圖 3.2-29 為將 5 萬筆評分資料依據 TARGET=1 機率遞減排序之程式碼，圖 3.2-30 為執行完排序後，取前 10000 名之程式碼。

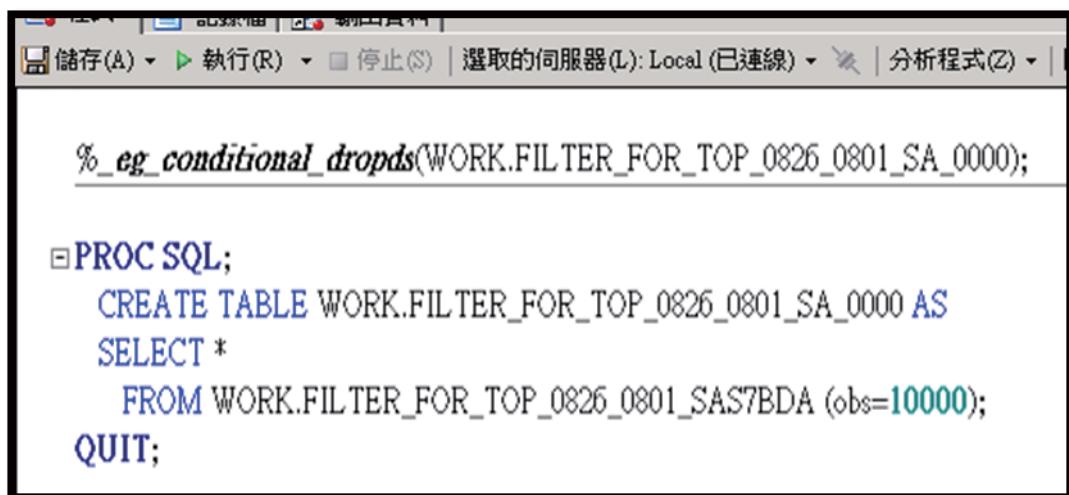
```

%_eg_conditional_dropds(WORK.FILTER_FOR_TOP_0826_0801_SAS7BDA);

PROC SQL;
  CREATE TABLE WORK.FILTER_FOR_TOP_0826_0801_SAS7BDA AS
  SELECT t1.ID, t1.P_TARGET1
    FROM EC100048.top_0826_0801 t1
   ORDER BY t1.P_TARGET1 DESC;
QUIT;

```

圖 3.2-29:排序之程式碼



The screenshot shows the SAS Enterprise Guide (EG) interface. The menu bar at the top includes '储存(A)', '执行(R)', '停止(S)', '选取的伺服器(L): Local (已连线)', '分析程式(Z)...', and other options. The main window displays the following PROC SQL code:

```
%_eg_conditional_dropds(WORK.FILTER_FOR_TOP_0826_0801_SA_0000);

PROC SQL;
  CREATE TABLE WORK.FILTER_FOR_TOP_0826_0801_SA_0000 AS
  SELECT *
    FROM WORK.FILTER_FOR_TOP_0826_0801_SAS7BDA (obs=10000);
QUIT;
```

圖 3.2-30:取前 10000 名之程式碼

## 四、研究結果

### 4.1 模型過度配置的觀察與處理

在找出最佳模組前，本研究透過 SAS Enterprise Guide 進行資料表整合，並整理出 3 筆資料集，分別為 SAS\_0731、SAS\_0801 以及 SAS\_0803。在初步建立 SAS\_0731 版本並執行預測後，本小組經反覆討論後，認為透過類神經模型的特性，輸入層若增加也許將能提升整體之準確率。故本小組將資料表欄位數再增加，使資料表之欄位內容更仔細，最後建立出 SAS\_0801 資料表而欄位數也從 54 增至 78，接著繼續經過變數觀察，在延伸為 108 個欄位數，此表為 SAS\_0803 版本，此為本研究使用原始資料集中欄位數最多的版本，如表 4.1-1 所示：

DATA SET		SAS_0731	SAS_0801	SAS_0803
衍生變數+target 欄位數		54	78	108
欄位相同處		顧客基本資料、預測目標變數		
欄位相異處	行內行為資料	將月份整合為 <u>半年</u> 資料	將月份整合為 <u>半年</u> 資料	將每個月份 <u>分開呈現</u> (即是從原始資料中，由列轉為欄)
	行內行為資料	將月份整合為 <u>半年</u> 資料	將月份整合為 <u>半年</u> 資料	將每個月份 <u>分開呈現</u> (即是從原始資料中，由列轉為欄)
	產品持有資料	無表示是否持有產品之 24 欄位	有表示是否持有產品之 24 欄位	有表示是否持有產品之 24 欄位

表 4.1-1: 三筆資料集之比較

經過模型評估後，運用 Model Comparison 節點進行三種資料表所建置之模型比較，發現衍生變數之欄位數愈多並不表示最後結果會愈符合期望，下一節將另行說明。

### 4.2 資料集與模型比較與結果

本研究整理出之 3 筆資料集 SAS\_0731、SAS\_0801 以及 SAS\_0803，皆運用如文章前述相同之模組與方法，各資料表所執行結果之 TRAIN 累積增益數據比較如表 4.2-1 所示：

MODEL\DATA SET	SAS_0731	SAS_0801	SAS_0803
Regression (迴歸)	3.36	4.38	3.33
Decision Tree-1 (決策樹-手動自建樹)	<b>4.09</b>	4.09	<b>4.41</b>
Neural Network (類神經網路)	3.9	<b>4.63</b>	4.01
Decision Tree (決策樹-自動樹)	1	1	1
SVM	2.49	3.31	2.33

表 4.2-1：累積增益數據比較表

結果顯示，使用 SAS\_0801 資料集，雖為 78 個變數，既使較 SAS\_0803 資料集之 108 個變數少，但預測分析卻擁有最佳結果，而 SAS\_0801 資料集使用 Neural Network (類神經) 模型累積增益值為 4.63，為最佳結果。

將此三種版本資料集的最佳模型取出，分別為 SAS\_0731 的 Decision Tree-1(決策樹-手動自建樹)、SAS\_0801 的 Neural Network (類神經網路) 與 SAS\_0803 的 Decision Tree-1(決策樹-手動自建樹)。並使用 MODEL COMPARISON 節點進行模型比較，如圖 4.2-1 所示，此圖為各資料表的最佳模型 ROC Chart 結果：

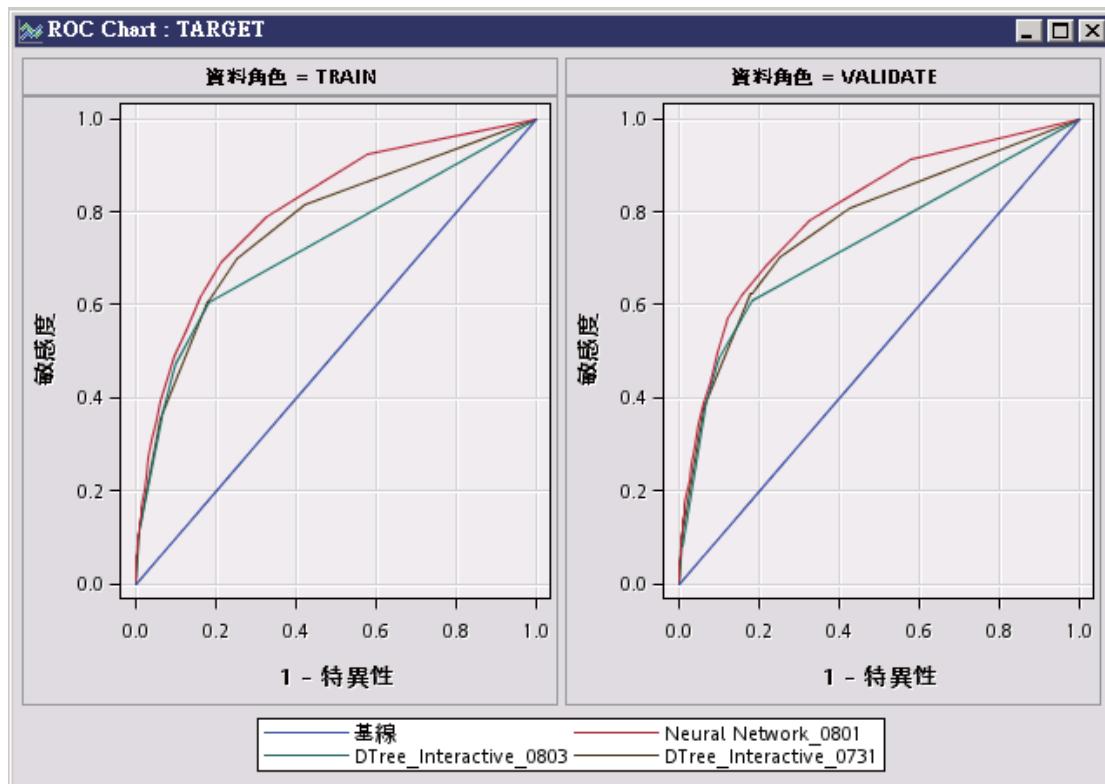


圖 4.2-1：各資料表的最佳模型 ROC Chart 結果

結果顯示，Neural Network\_0801 即 SAS\_0801 資料集之 Neural Network 類神經模型 ROC Curve 為最理想之結果。由以上數據與圖表所示，因此本團隊最後決定使用 SAS\_0801 資料集的 Neural Network 類神經模型為最後的

決策模型。本團隊最後所提出的 10000 筆建議銷售名單成交率的預測機率範圍最高至最低介於約 **0.70 至 0.05** 之間。以下為提出的 10000 筆建議銷售名單的部分截圖，如圖 4.2-2 所示：

ID	P_TARGET1	ID	P_TARGET1
1	0.7053771961	9971	0.0575286749
2	0.6980060134	9972	0.0575045617
3	0.6726039289	9973	0.0574986939
4	0.648484229	9974	0.0574890467
5	0.6203811736	9975	0.0574873915
6	0.5568213432	9976	0.0574863974
7	0.5440763532	9977	0.0574751709
8	0.5156742867	9978	0.0574666323
9	0.5117159738	9979	0.0574645082
10	0.4976062237	9980	0.0574599185
11	0.4799879992	9981	0.0574592712
12	0.4721283429	9982	0.0574545059
13	0.4697247701	9983	0.0574533677
14	0.467232051	9984	0.0574441819
15	0.4614766743	9985	0.0574320314
16	0.4572389467	9986	0.0574297441
17	0.4562581706	9987	0.0574269401
18	0.4511461127	9988	0.0574208742
19	0.4485929763	9989	0.0574180955
20	0.4456979614	9990	0.057412654
21	0.4453308057	9991	0.057396815
22	0.444922778	9992	0.0573933786
23	0.4433068531	9993	0.0573880218
24	0.4307527348	9994	0.0573867929
25	0.4302222021	9995	0.0573862046
26	0.4275284777	9996	0.0573814322
27	0.4237514226	9997	0.0573729989
28	0.4191969568	9998	0.057357776
29	0.4184198908	9999	0.0573557408
30	0.4164268489	10000	0.0573526051
31	0.4099406948	10001	0.0573505897
32	0.4016223128	10002	0.0573464375

圖 4.2-2：提出 10000 筆建議銷售名單部分截圖

### 4.3 結果驗證

5 萬筆評分資料經運算預測輸出後，為了檢視其準確度，本小組運用此模組中之最佳模型-類神經模型，運用原始資料經類神經模型訓練後之預測 TARGET 為 1 之結果資料，與經預測後之 5 萬筆評分資料的前一萬筆資料做比較，如圖 4.3-1 所示，發現兩者之比率雷同，表示此模型預測結果準確率之可信度相當大。

<p>類神經TRAIN後的預測前10000筆 約0.699~0.046</p>		VS	<p>評分資料之前10000筆結果 約0.70~0.05</p>	
<p>表格前幾筆</p>		<p>表格後幾筆 (接近10000筆)</p>		
Predicted: TARGET=1		Predicted: TARGET=1		
0.6991466882363028		0.046524264527180864		<b>P_TARGET1</b>
0.6820556102795244		0.04652349188721981		0.0575286749
0.6130868717050126		0.04651269903703303		0.0575045617
0.5996865294557693		0.04651122910199744		0.0574986939
0.58847774524309		0.046506084952485274		0.0574890467
0.5867333122366386		0.046502793632550506		0.0574873915
0.5723963007886225		0.04650199223395472		0.0574863974
0.568869269483789		0.04649787131836899		0.0574751709
0.5683471953217021		0.046494418654646506		0.0574666323
0.5637847749925227		0.046485961261503755		0.0574645082
0.5448052251946196		0.04648523482422611		0.0574599185
0.5208152432423635		0.046460262608858376		0.0574592712
0.5161686896942801		0.04645653195654572		0.0574545059
0.49785395227881557		0.046455414806799454		0.0574533677
0.49559672325526327		0.04645495734835642		0.0574441819
0.4952159736928117		0.04645491503182343		0.0574320314
0.48381632742148406		0.04645473799184959		0.0574297441
0.4714063187687159		0.046453572857130136		0.0574269401
0.46423435075285824		0.046448625879094575		0.0574208742
0.46423435075285824		0.046443113573473124		0.0574180955
0.4565804698188759		0.04643837497795422		0.057412654
0.4535537640260419		0.04643481030152676		0.057396815
0.45267060298644124		0.046433895172033786		0.0573933786
0.4501221598139194		0.046433128177941624		0.0573880218
0.44517521412275063		0.04642501577323156		0.0573867929
0.4350308140539114		0.04642369059286354		0.0573862046
0.42859089700742414		0.04641519715570295		0.0573814322
		0.04641000411711872		0.0573729989
		0.0464064612820961		0.057357776

圖 4.3-1:原始資料經類神經模型訓練後之預測 TARGET 為 1 之結果資料與經預

#### 4.4 結論

- 一、 經衍生變數評估與比較後，最終運用本小組整合衍伸後之 SAS\_0801 資料集進行預測，最後所提出 10000 筆建議銷售名單，其成交率的範圍最高至最低介於約 0.70 至 0.05 之間。
- 二、 經比較評估各資料集與使用模型所產生之數據與圖表結果，本團隊最後決定使用 SAS\_0801 資料集的 Neural Network 類神經模型為最後的決策模型。
- 三、 經反覆測試發現衍生變數愈多並不表示結果會愈符合預期，必須經過詳細評估與比較才能得到最佳結果，且最適模型的種類也可能會因此有所改變。

### 五、 參考文獻

- T. Hastie, *et al.*, "The element of statistical learning," *Springer-Verlag*, 2001.
- 陳亭羽 and 賀千盈, "應用決策術探討適用於電子行銷市場之區隔基礎," 管理科學研究, vol. 3, pp. 1-25, 2006.
- 郭一聰, "一用決策術與類神經網路於應收帳款之呆帳預警模式研究," 2005.
- 沈偉立, "結合迴歸決策術語類神經網路於時間序列資料之預測-以台股指數期貨與國民生產毛額為例," January, 2007.
- 柯坤志, *et al.*, "支撐向量迴歸分析在股票價格上之預測與軟體開發," presented at the 資訊科技國際研討會, 台中, 2009.
- C. Judy, *et al.*, "衛生統計資料倉儲系統第二階段資料採礦建置結案報告書 v1.0," 2004.

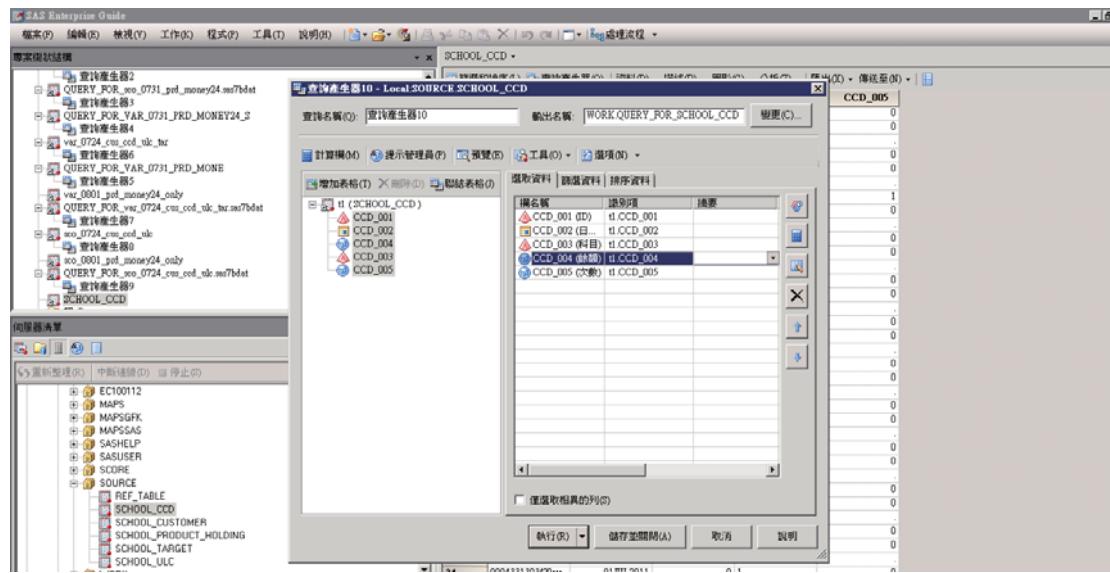
## A1: 附錄一

### SAS Enterprise Guide (EG)操作流程 Step by Step

A1: 附錄一 SAS Enterprise Guide (EG)操作流程 Step by Step	
A1-1 SCHOOL_CCD 和 SCHOOL_ULC 之欄位新建擴增 .....	A1-1
A1-2 SCHOOL_CCDN、SCHOOL_ULCN、SCHOOL_CUSTOMER 和 SCHOOL_TARGE 做連結 .....	A1-4
A1-3 SCHOOL_PRODUCT_HOKDING 之欄位新建擴增 .....	A1-9
A1-4 SCHOOL_CUS_CCDULC_TAR、SCHOOL_PRO24N 及 SCHOOL_PRO24B 做連結 .....	A1-14

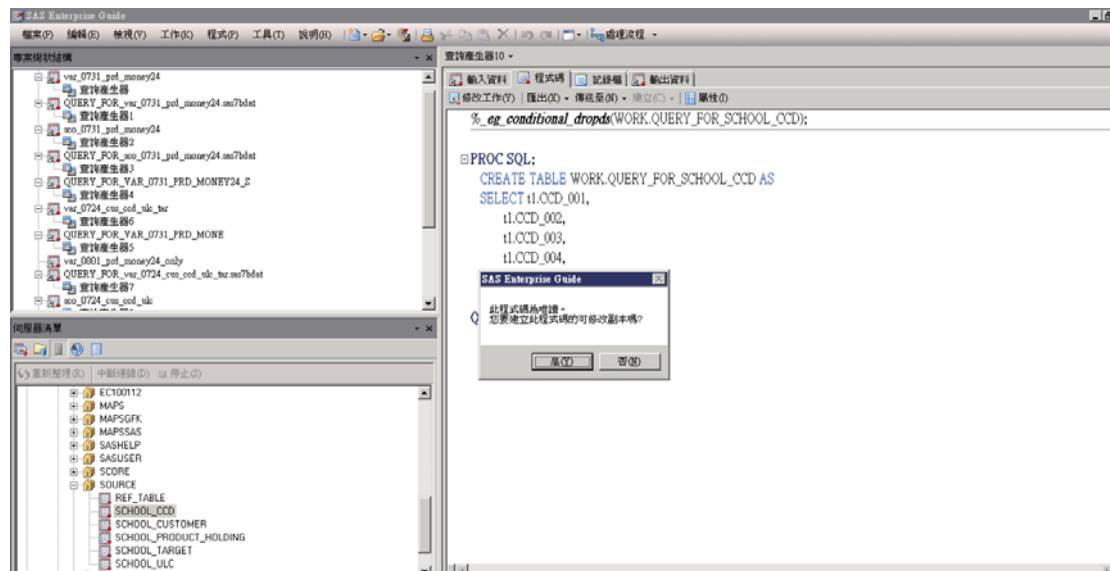
## A1: 附錄一 SAS Enterprise Guide (EG)操作流程 Step by Step

### 1 SCHOOL\_CCD 和 SCHOOL\_ULC 之欄位新建擴增



#### STEP 1:

首先針對 SCHOOL\_CCD 之檔案，將它從左下方的 SOURCE 的資料夾拉出，接著對此檔案點選右鍵，選擇[查詢產生器]，接著將左方 5 個變數拉入右邊至此畫面在按下執行。



#### STEP 2:

按下執行後，選擇視窗中間的[程式碼]，接著會進入程式碼的畫面，點下程式碼進行修改，輸入的同時它會問你是否要建立程式碼之副本，選擇[是]。

```

%_eg_conditional_drops(WORK.QUERY_FOR_SCHOOL_CCD);

PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_SCHOOL_CCD AS
  SELECT SUM(t1.CCD_004),SUM(t1.CCD_005),
  t1.CCD_001
  FROM SOURCE.SCHOOL_CCD t1
  WHERE t1.CCD_003=1
  Group by t1.CCD_001;
QUIT;

```

### STEP 3:

進行程式碼的修改，根據上圖將程式碼寫入。因每一個 ID 都有 18 筆的資料，18 筆的資料包含 3 種不同的帳號狀態以及 6 個月分的組合，在這邊將其依照 3 種帳戶的狀態擴增 3 為新的 3 欄，且將 6 個月合併成為同一欄。

	_TEMG001	_TEMG002	CCD_001
1	7241	1	000493130126***
2	6	0	000493130140***
3	4117	7	000493130105***
4	22121	23	000493130117***
5	11734	24	000493130117***
6	37553	10	000493130129***
7	15300	4	000493130121***
8	0	0	000493130120***
9	23568	8	000493130130***
10	5797	3	000493130102***
11	12317	6	000493130127***
12	17872	13	000493130140***
13	181627	60	000493130144***
14	16395	22	000493130146***
15	8830	4	000493130147***
16	19293	7	000493130126***
17	62473	17	000493130152***
18	11646	11	000493130161***
19	0	0	000493130167***
20	39678	6	000493130162***
21	28524	9	000493130170***
22	64169	39	000493130120***
23	3957	7	000493130172***
24	11968	6	000493130173***
25	21297	16	000493130181***
26	80298	15	000493130129***
27	44969	11	000493130169***
28	104099	11	000493130192***
29	29922	20	000493130192***
30	12421	13	000493131302***
31	40207	24	000493131322***
32	38506	15	000493131322***
33	77053	12	000493131319***

### STEP 4:

此圖為輸出之結果，第一欄是 CCD\_003 的總金額，第二欄是 CCD\_003 的總餘額。以此類推其餘的兩種帳戶狀態來建立另兩筆資料集。



### STEP 5:

利用查詢產生器將帳戶的三種不同狀態(一般消費、預借現金以及上期未繳金額之循環) 合併成同一筆資料，打開[查詢產生器]後左上方有個[增加表格]，點開後在點左上方的[增加表格]，將剛剛新建出的三筆資料做連結，在 ID 的欄位上點右鍵，選擇[連結至下列項目到 T2 以及 T3 的連結]再按下[確定]後。

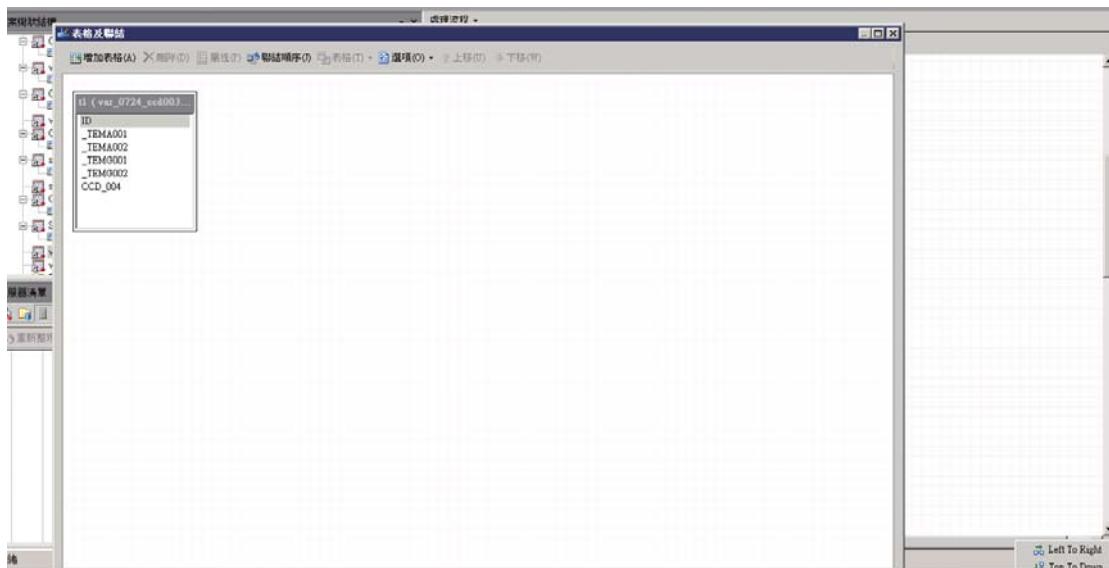
ID	_ITEM001	_ITEM002	_ITEM003	CCD_004
1	000433130329...	7241	1	0
2	000433130420...	0	0	0
3	000433130305...	4117	7	0
4	000433130317...	21212	23	0
5	000433130317...	11734	24	0
6	000433130309...	37553	10	0
7	000433130321...	15300	4	0
8	000433130320...	0	0	0
9	000433130306...	23560	8	0
10	000433130312...	5707	3	0
11	000433130327...	12317	6	0
12	000433130304...	17872	13	0
13	000433130304...	181627	60	0
14	000433130346...	16395	22	0
15	000433130347...	8830	4	10000
16	000433130306...	19293	7	0
17	000433130325...	62473	17	0
18	000433130301...	11646	11	0
19	000433130307...	0	0	60000
20	000433130305...	39670	6	0
21	000433130304...	26524	9	0
22	000433130306...	64169	39	0
23	000433130312...	3957	7	0
24	000433130370...	11968	6	0
25	000433130313...	21297	16	0
26	000433130326...	88288	15	0
27	000433130309...	44969	11	0
28	000433130300...	184099	11	0
29	000433130302...	29922	20	0
30	000433131100...	12421	13	0
31	000433131100...	40207	24	0
32	000433131100...	38506	15	0
33	000433131104...	27953	12	0
34	000433131128...	58189	14	0

### STEP 6:

此圖為 SCHOOL\_CCD 三種帳戶狀態之金額及次數輸出結果，因上期未繳金額此狀態因只要看最後一次，所以次數部分無須多做欄位。

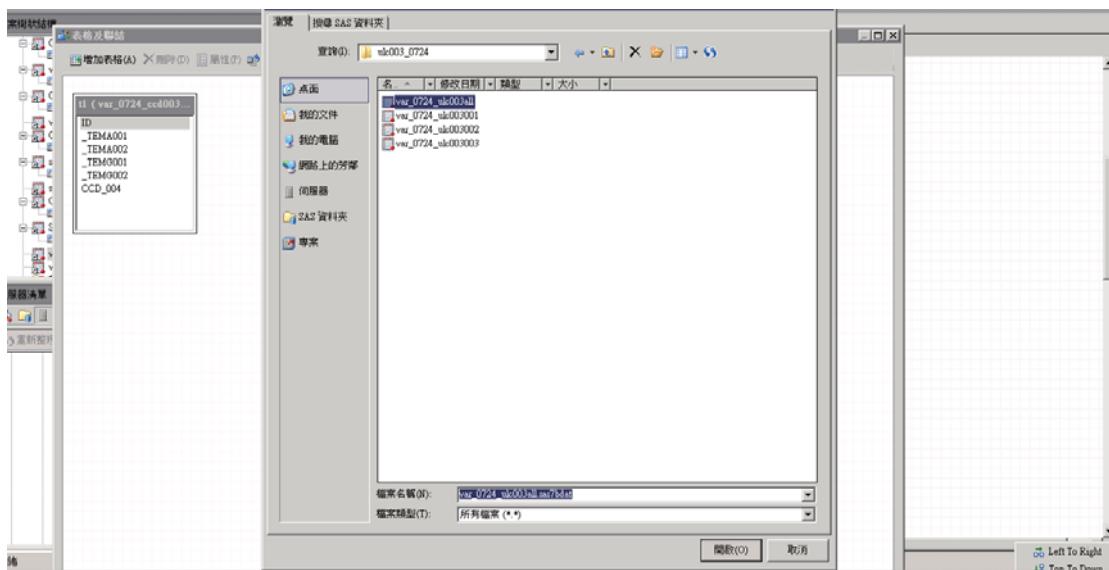
SCHOOL\_ULC 做法跟 SCHOOL\_CCD 相同，但 SCHOOL\_ULC 因為行外資料，故無次數欄位，所以在做法上可將次數排除。

## 2 SCHOOL\_CCDN、SCHOOL\_ULCN、 SCHOOL\_CUSTOMER 和 SCHOOL\_TARGE 做連結



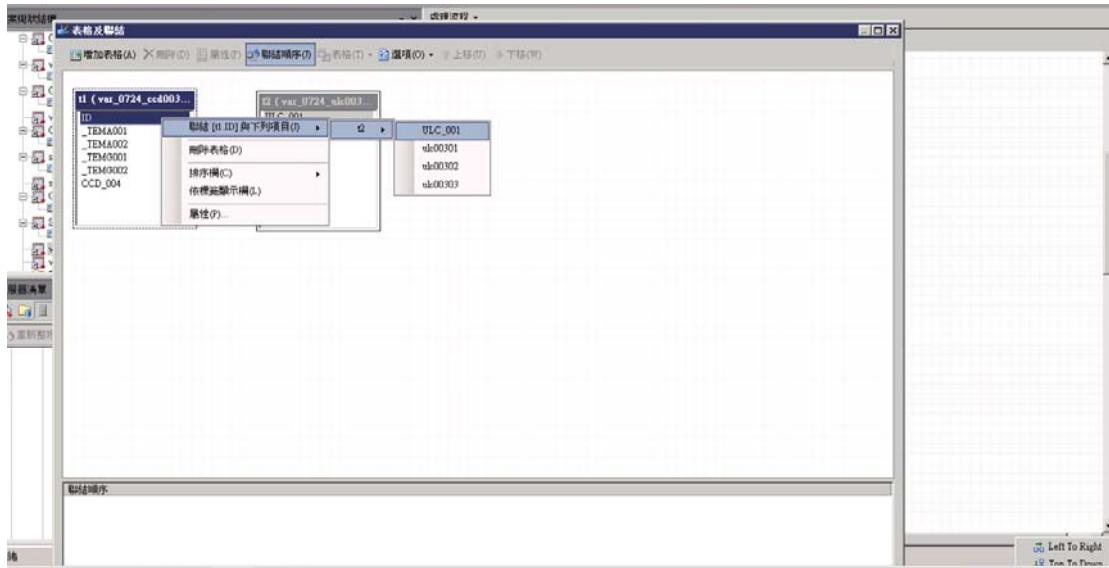
STEP 1:

將 SCHOOL\_CCD 及 SCHOOL\_ULC 擴增之兩種新資料集做連結，這邊將新資料集另外命為 SCHOOL\_CCDN 及 SCHOOL\_ULCN。將滑鼠指到 SCHOOL\_CCDN 且按下右鍵，選擇[查詢產生器]。



STEP 2:

打開視窗選擇左上方的[增加表格]，接著開起視窗在選擇左上方的[增加表格]，選擇要做連結之資料集(SCHOOL\_ULCN)。



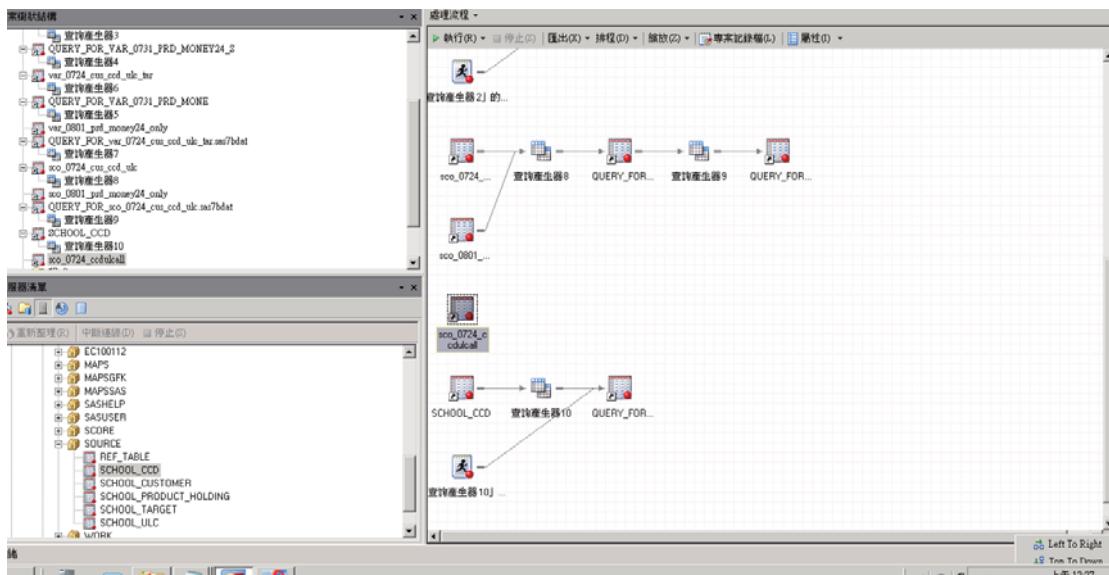
### STEP 3:

在 T1 的 ID 上按右鍵，選擇[連結 T1.ID 與下列項目]，連結 T2 的 ID。

	ID(ID)	ccd00301	ccd00302	ccd00303	ccd00304	ccd00305	vlc00301	vlc00302
1	000493130316***	224275	16	0	0	0	591200	16
2	000493230302***	44355	41	0	0	0	39449	
3	000443111115***	90509	31	0	0	0	52769	
4	000493130301***	22456	1	0	0	0	33993	
5	000493130322***	20598	17	50000	1	0	20753	50
6	000493130300***	28822	16	0	0	0	12990	33168
7	000493130321***	203100	110	0	0	0	39966	
8	000493130318***	15772	49	0	0	0	10437	18828
9	000493130322***	40565	25	0	0	0	36978	
10	000493130325***	2077	4	0	0	0	0	-2189
11	000493130302***	13143	17	0	0	0	0	42860
12	000493130304***	0	0	0	0	0	0	-15038
13	000493130302***	0	0	0	0	0	0	36976
14	000493130304***	98257	69	0	0	0	0	111416
15	000493130304***	108943	70	0	0	0	0	103443
16	000493130307***	57015	18	0	0	0	0	65981
17	000493130321***	30604	28	0	0	0	0	83866
18	000493130300***	34605	11	0	0	0	0	16678
19	000493130314***	88435	25	0	0	0	26000	89607
20	000493130305***	135916	76	0	0	0	0	29782
21	000493130307***	0	0	0	0	0	34260	4415
22	000493130311***	10029	10	0	0	0	0	4315
23	000493130303***	0	0	0	0	0	0	0
24	000493130307***	05454	23	0	0	0	0	31029
25	000493130307***	60654	22	0	0	0	0	84544
26	000493130325***	12475	1	0	0	0	0	63878
27	000493130304***	43168	34	0	0	0	8300	50379
28	000493130303***	14011	13	0	0	0	0	11211
29	000493130303***	21126	20	0	0	0	12990	50771
30	000493130303***	23330	25	0	0	0	6318	48489
31	000493130300***	161475	111	0	0	0	0	335594
32	000493130302***	0	0	410000	1	0	91532	27
33	000493130300***	31599	51	6000	1	49300	43056	6

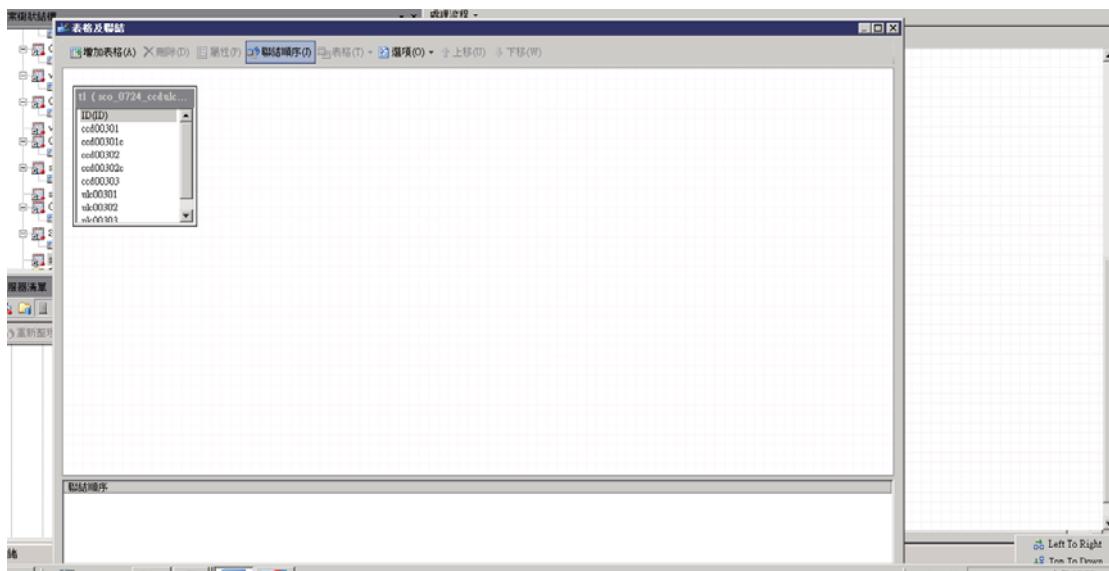
### STEP 4:

以下為 SCHOOL\_CCDN 及 SCHOOL\_ULCN 連結完之輸出結果，將它命名為 SCHOOL\_CCDULC。



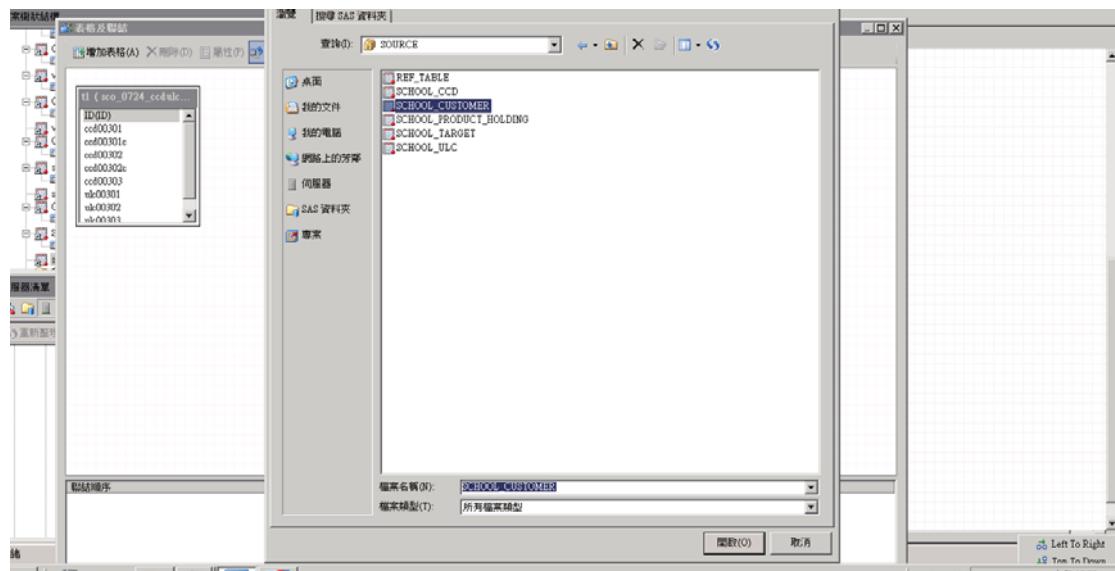
### STEP 5:

對 SCHOOL\_CCDULC 此資料集與 SCHOOL\_CUSTOMER 和 SCHOOL\_TARGET 做連結。在 SCHOOL\_CCDULC 資料集上點滑鼠右鍵，選擇[查詢產生器]，打開視窗後選擇左上方的[增加表格]。



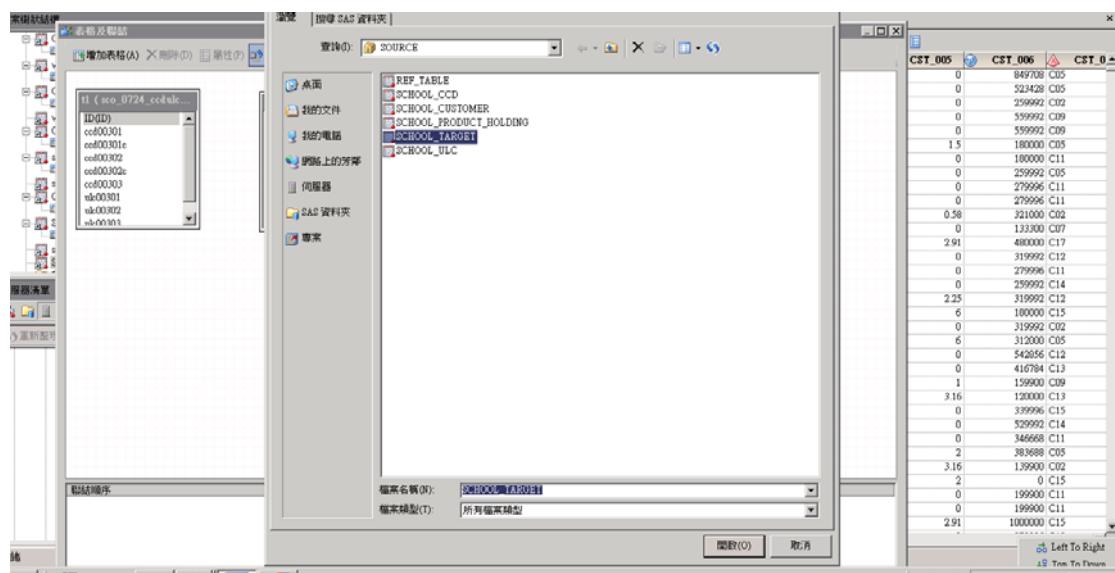
### STEP 6:

打開視窗後選擇左上方之[增加表格]。



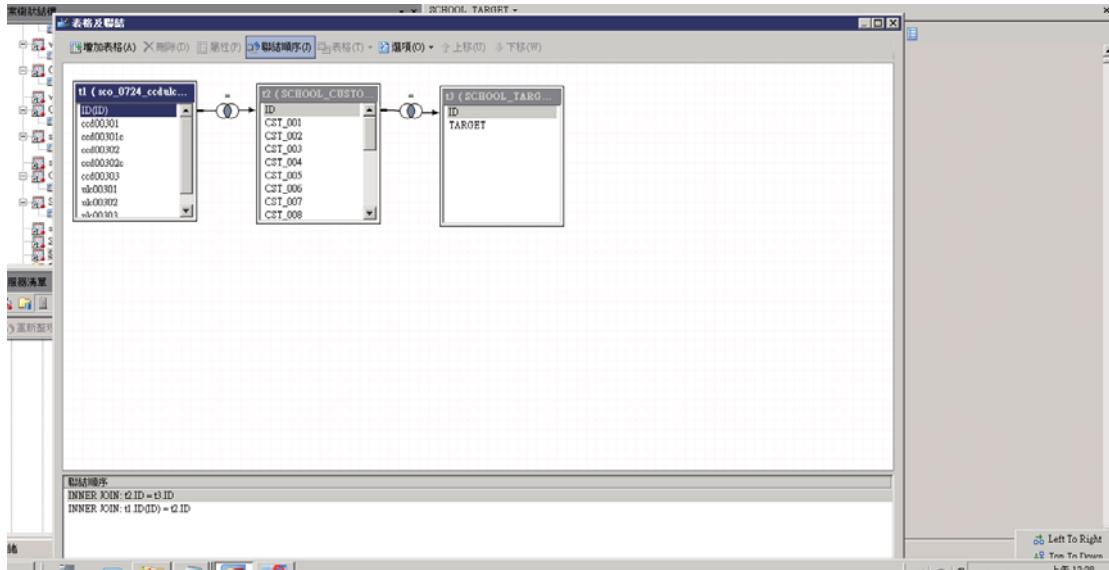
### STEP 7:

選擇欲加入的資料集，SCHOOL\_CUSTOMER。



### STEP 8:

選擇資料集 SCHOOL\_TARGE。



### STEP 9:

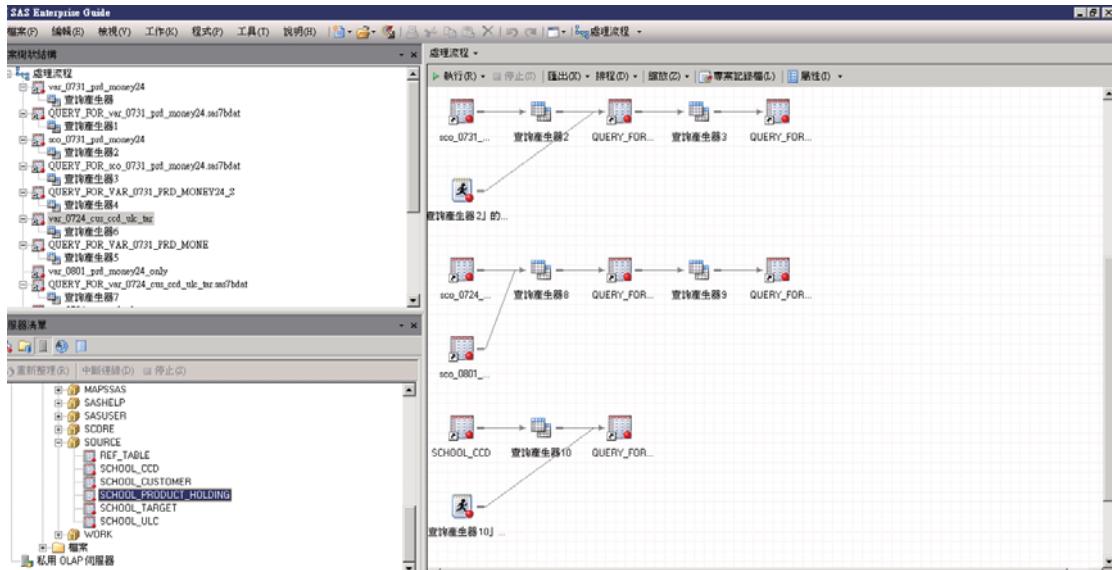
將三個資料集做聯結，[T1.ID]對上[T2.ID]，[T2.ID]對上[T3.ID]。

	ID	CST_001	CST_002	CST_003	CST_004	CST_005	CST_006	CST_007
1	000433130329***	M	0.5	1	0	0	849708 C05	
2	000433130340***	M	0.8	1	0	0	523428 C05	
3	000469130310***	M	0.4	2	0	0	259992 C02	
4	000469130317***	M	0.5	2	0	0	559992 C09	
5	000469130317***	M	0.5	2	0	0	559992 C09	
6	000469130329***	M	0.5	2	1.5	0	180000 C05	
7	000469130321***	M	0.4	2	0	0	180000 C11	
8	000469130321***	M	0.5	1	0	0	259992 C05	
9	000469130300***	M	0.4	2	0	0	279996 C11	
10	000469130317***	M	0.4	2	0	0	279996 C11	
11	000469130327***	M	0.3	2	0.58	0	321000 C02	
12	000469130340***	M	0.8	1	0	0	133300 C07	
13	000469130340***	M	0.1	2	2.91	0	480000 C17	
14	000469130340***	M	0.3	2	0	0	319992 C12	
15	000469130342***	M	0.4	2	0	0	279996 C11	
16	000469130306***	M	0.4	2	0	0	259992 C14	
17	000469130322***	M	0.3	2	2.25	0	319992 C12	
18	000469130361***	M	0.5	2	6	0	180000 C15	
19	000469130367***	M	0.4	2	0	0	319992 C02	
20	000469130326***	M	0.4	2	6	0	312000 C05	
21	000469130300***	M	0.5	2	0	0	542856 C12	
22	000469130300***	M	0.3	2	0	0	416784 C13	
23	000469130372***	M	0.3	2	1	0	159990 C09	
24	000469130373***	M	0.3	2	3.16	0	120000 C13	
25	000469130381***	M	0.3	2	0	0	339996 C15	
26	000469130320***	M	0.4	1	0	0	559992 C14	
27	000469130360***	M	0.4	2	0	0	346668 C11	
28	000469130320***	M	0.3	2	2	0	303688 C05	
29	000469130322***	M	0.2	2	3.16	0	139990 C02	
30	000469131300***	M	0.3	2	2	0	0 C15	
31	000469131322***	M	0.4	2	0	0	199990 C11	
32	000469131322***	M	0.4	2	0	0	199990 C11	
33	000469131306***	M	0.5	1	2.91	0	1000000 C15	

### STEP 10:

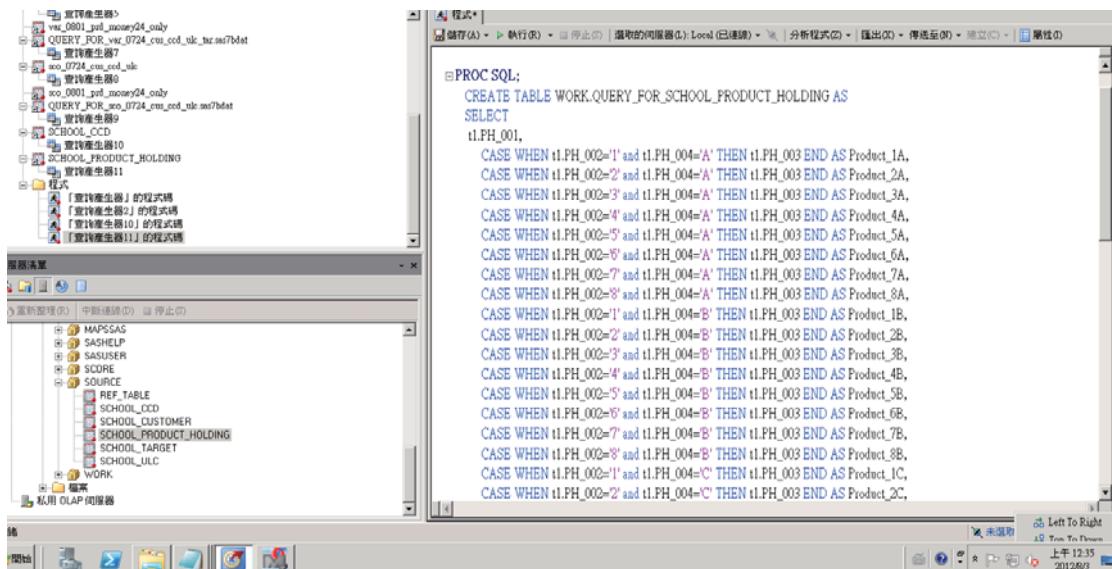
此圖為輸出之結果，將此圖命名為 SCHOOL\_CUS\_CCDULC\_TAR。

### 3 SCHOOL\_PRODUCT\_HOKDING 之欄位新建擴增



#### STEP 1:

選擇 SCHOOL\_PRODUCT\_HOKDING 之檔案，要對此資料去做寫程式碼之動作。



#### STEP 2:

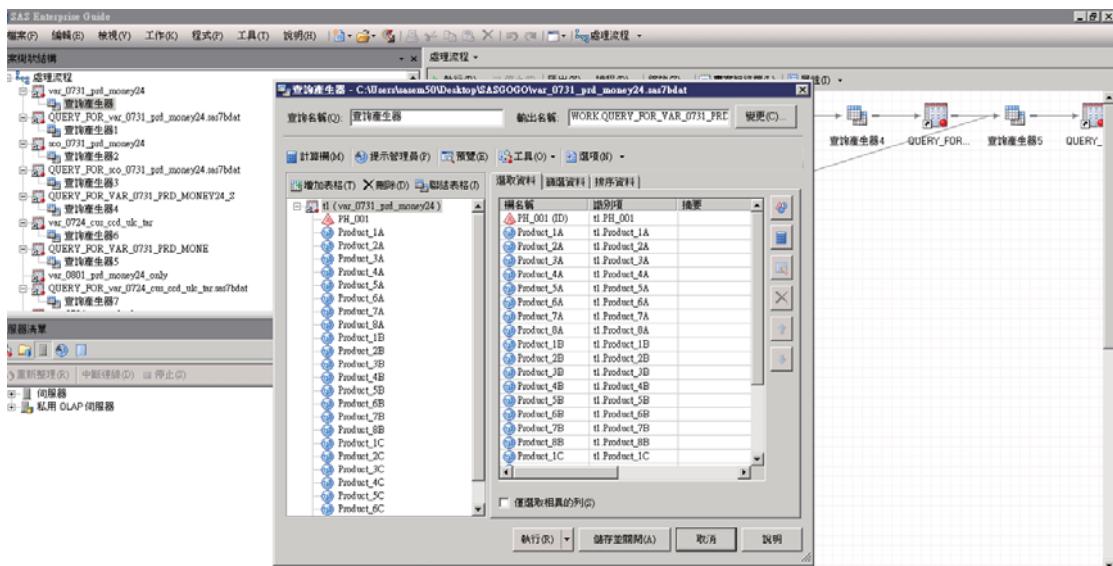
對此資料點擊右鍵，選擇[查詢產生器]，將視窗左邊的資料都拉入右方，接著按下[確定]。到了輸出畫面後，選擇視窗上方的[程式]，打開寫碼的視窗。程式法的部分依照上列方式輸入。

[CASE WHEN t1.PH\_002='1' and t1.PH\_004='A' THEN t1.PH\_003 END AS Product\_1A]

此行程式碼中，PH\_002 為產品分類之編號，產品編號有八種，PH\_003 為帳戶之餘額，PH\_004 為帳戶狀態，帳戶狀態有三種。是將 PH\_002 的八種不同分類與 PH\_004 的三種帳戶狀態做組合，所以會產生出 24 種新的欄位。

### STEP 3:

此圖為輸出結果，為 24 個欄位。欄位用阿拉伯數字及英文符號做欄位之命名。  
此資料集命名為 SCHOOL\_PRO24。



### STEP 4:

因 SCHOOL\_PRODUCT\_HOLDING 此資料集比較特別,一個 ID 可能有多項產品的資料但也可能沒有購買商品，所以需要將此新資料集進行整合。將各欄的資料進行加總，讓可能會重複的產品之 ID 進行加總。

在此資料集上點擊右鍵，選擇[查詢產生器]，在視窗中將左方所有資料拖曳至右方，按下[執行]。接著在輸出結果的視窗上選擇[程式]，進行寫碼。

### STEP 5:

此圖為程式碼的部分，將 24 個欄位的進行加總

[ $\text{SUM}(\text{t1.Product\_1A})$ ]

此行程式碼是將欄位 Product\_1A 做加總，以 PH\_001 為 GROUP。

### STEP 6:

此圖為加總程式碼。

SAS Enterprise Guide

處理流程 - QUERY\_FOR\_var\_0731\_pf1\_money24.m7bdst

查詢結果

	PH_001	_TEMG001	_TEMG002	_TEMG003	_TEMG004	_TEMG005	_TEMG006	_TEMG007
1	0004931.0311***	3313	44980					
2	0004931.0311***	3313	44980					
3	0004931.0320***	2830		206977				
4	0004931.0330***							
5	0004931.0332***	19551						
6	0004931.0340***	120						
7	0004931.0340***	419755						
8	0004931.0340***	0						
9	0004931.0342***	44980						
10	0004931.0350***	12331						
11	0004931.0361***	16876		230004				
12	0004931.0361***	0						
13	0004931.0365***	3657979.9						
14	0004931.0370***	246						
15	0004931.0372***	7923		171920				
16	0004931.0381***	34653						
17	0004931.0395***	10629						
18	0004931.0402***	1360						
19	0004931.0404***	115145						
20	0004931.0412***							
21	0004931.0413***	861		254977				
22	0004931.0413***	19551						
23	0004931.0420***	10629						
24	0004931.0420***	1154						
25	0004932.0311***	3657979.9	44980	190000	3658000			
26	0004932.0310***	305		171920				
27	0004932.0312***	25910						
28	0004932.0316***	0						
29	0004932.0322***	57992						
30	0004932.0339***							
31	0004932.0340***	9382						
32	0004932.0340***	5651						
33	0004932.0340***	4819		171920				

### STEP 7:

此圖為將 24 個欄位加總之輸出結果。將此資料集命名為 SCHOOL\_PRO24N。

SAS Enterprise Guide

處理流程 - WORK.QUERY\_FOR\_VAR\_0731\_FRD\_MONE\_0000

查詢產生器4 - Local.WORK.QUERY\_FOR\_VAR\_0731\_FRD\_MONE\_0000

查詢名稱: 查詢產生器4

輸出名稱: WORK.QUERY\_FOR\_VAR\_0731\_FRD\_MONE\_0000

計算機(D) 提示管理員(P) 視窗(E) 工具(O) 選項(N)

增加表格(T) × 檢附(D) 計算表格(B)

擷取資料 | 擷取資料 | 資料序列 |

欄名表 | 欄別類 | 摘要 |

計算機

	詳細資料
Product_1_A	CASE WHEN t1.Product_1A > 0 THEN 1 ELSE 0.
Product_1_B	CASE WHEN t1.Product_1B > 0 THEN 1 ELSE 0.
Product_1_C	CASE WHEN t1.Product_1C > 0 THEN 1 ELSE 0.
Product_2_A	CASE WHEN t1.Product_2A > 0 THEN 1 ELSE 0.
Product_2_B	CASE WHEN t1.Product_2B > 0 THEN 1 ELSE 0.
Product_2_C	CASE WHEN t1.Product_2C > 0 THEN 1 ELSE 0.
Product_3_A	CASE WHEN t1.Product_3A > 0 THEN 1 ELSE 0.
Product_3_B	CASE WHEN t1.Product_3B > 0 THEN 1 ELSE 0.
Product_3_C	CASE WHEN t1.Product_3C > 0 THEN 1 ELSE 0.
Product_4_A	CASE WHEN t1.Product_4A > 0 THEN 1 ELSE 0.
Product_4_B	CASE WHEN t1.Product_4B > 0 THEN 1 ELSE 0.
Product_4_C	CASE WHEN t1.Product_4C > 0 THEN 1 ELSE 0.
Product_5_A	CASE WHEN t1.Product_5A > 0 THEN 1 ELSE 0.

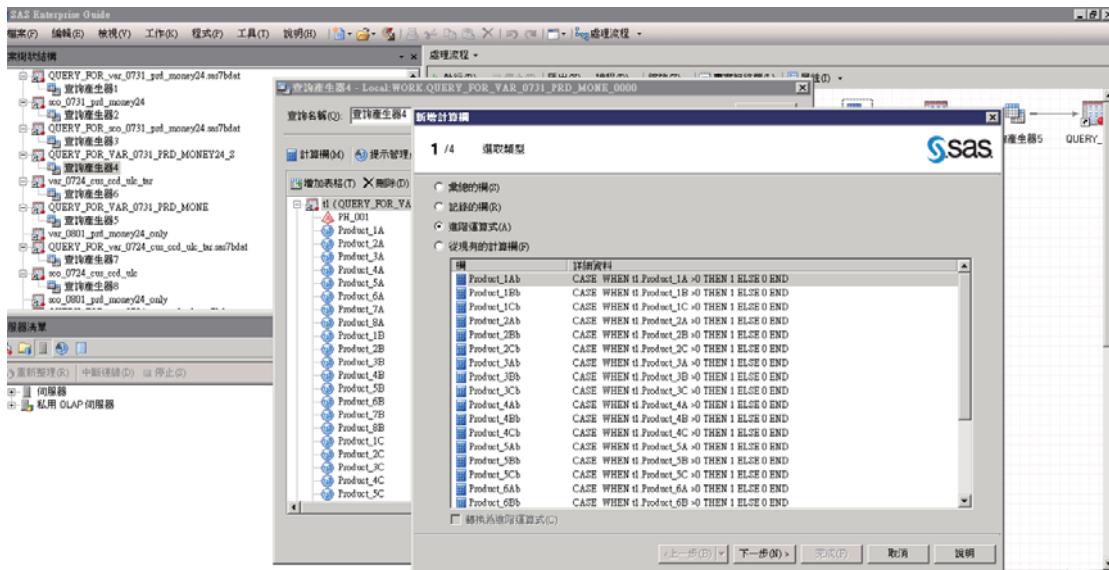
新增(N)... | 編輯(E)... | 刪除(D)... | 重新命名(R)... | 關閉 |

執行(F) | 儲存並關閉(A) | 取消 | 說明

### STEP 8:

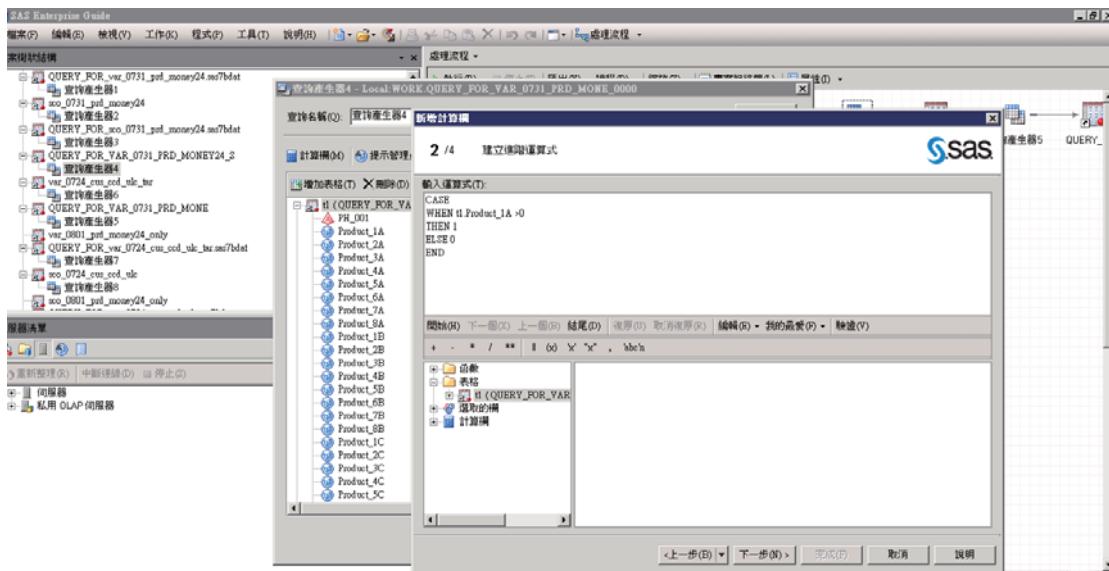
因加總過後，會有遺失值的產生，在此處我們須將新建出的 24 個欄位去做擴增，將有顯示數值的欄位顯示為 1，其他的為 0(包括遺失值)。

對 SCHOOL\_PRO24N 按右鍵，選擇[查詢產生器]後，在視窗中選擇[計算機]，接著選擇視窗中右方的[新增]。



### STEP 9:

進入此視窗後，選擇進階運算式且進行下一步。

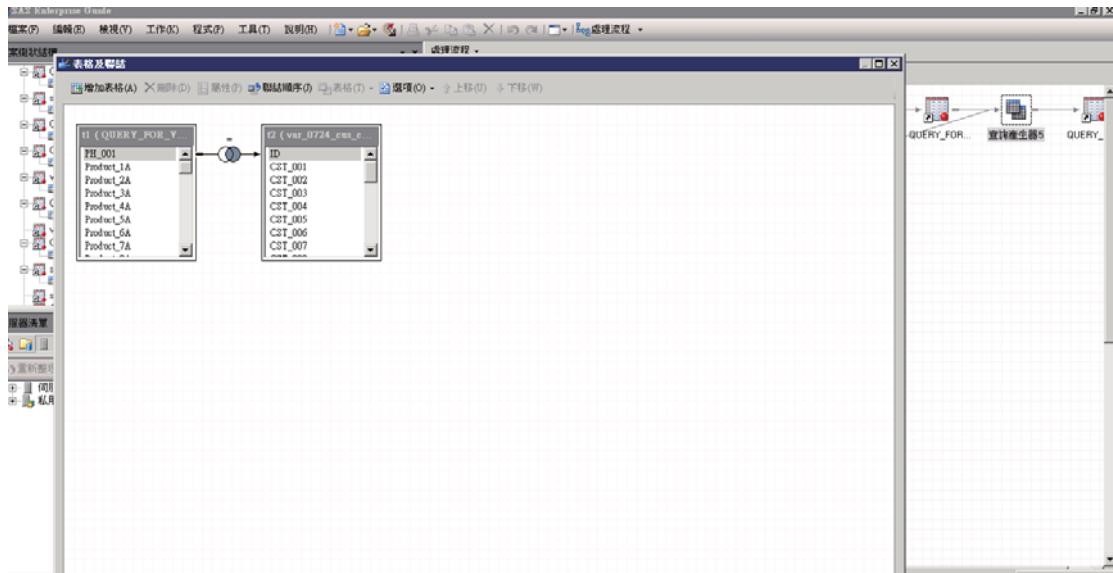


### STEP 10:

寫入運算式，可在左下方的函數資料夾使用已寫好的函數，我們所使用的函式 CASE。程式碼照此圖輸入即可，但欄位名稱須依照不同之欄位做更改。完成後按下[下一步]直至完成即可。此資料集命名為 SCHOOL\_PRO24B。

## 4 SCHOOL\_CUS\_CCDULC\_TAR、SCHOOL\_PRO24N 及

### SCHOOL\_PRO24B 做連結



#### STEP 1:

接著須將所有建立好之資料集做連結，將 SCHOOL\_CUS\_CCDULC\_TAR 、 SCHOOL\_PRO24N 及 SCHOOL\_PRO24B 做連結，在 SCHOOL\_CUS\_CCDULC\_TAR 上點擊右鍵，出現[查詢產生器]，接著在視窗左上方選擇[增加表格]，接著在進入視窗後選擇[增加表格]，接著選擇你所要之檔案做[連結]，[T1.ID]對上[T2.ID]，[T2.ID]對上[T3.ID]即可。

The screenshot shows the '查詢產生器5' (Query Generator 5) window. The left pane lists various query definitions and their execution status. The right pane displays a large table with 32 rows and 8 columns, representing the joined data from multiple sources. The columns are labeled: ID, CST\_001, CST\_002, CST\_003, CST\_004, CST\_005, CST\_006, and CST\_007.

ID	CST_001	CST_002	CST_003	CST_004	CST_005	CST_006	CST_007
000493130332*** M	0.5	1		0	849708	C05	
000493130342*** M	0.8	1		0	523428	C05	
000493130352*** M	0.4	2		0	259992	C02	
000493130359*** M	0.5	2		0	559992	C09	
000493130317*** M	0.5	2		0	559992	C09	
000493130320*** M	0.5	2		1.5	180000	C05	
000493130321*** M	0.4	2		0	180000	C11	
000493130328*** M	0.5	1		0	259992	C05	
000493130336*** M	0.4	2		0	279996	C11	
000493130339*** M	0.4	2		0	279996	C11	
000493130340*** M	0.3	2		0.58	321000	C02	
000493130341*** M	0.8	1		0	133300	C07	
000493130346*** M	0.1	2		2.91	480000	C17	
000493130347*** M	0.3	2		0	319992	C12	
000493130347*** M	0.4	2		0	279996	C11	
000493130356*** M	0.4	2		0	259992	C14	
000493130352*** M	0.3	2		2.25	319992	C12	
000493130361*** M	0.5	2		6	180000	C15	
000493130361*** M	0.4	2		0	319992	C02	
000493130365*** M	0.4	2		6	312000	C05	
000493130366*** M	0.5	2		0	542856	C12	
000493130366*** M	0.3	2		0	416784	C13	
000493130371*** M	0.3	2		1	159900	C09	
000493130378*** M	0.3	2		3.16	120000	C13	
000493130381*** M	0.3	2		0	339996	C15	
000493130382*** M	0.4	1		0	529992	C14	
000493130390*** M	0.4	2		0	346660	C11	
000493130392*** M	0.3	2		2	386888	C05	
000493130392*** M	0.2	2		3.16	139900	C02	
000493131302*** M	0.3	2		2	0	C15	
000493131302*** M	0.4	2		0	199900	C11	
000493131302*** M	0.4	2		0	199900	C11	

#### STEP 2:

此圖為最終輸出之結果。欄位共有 78 欄。

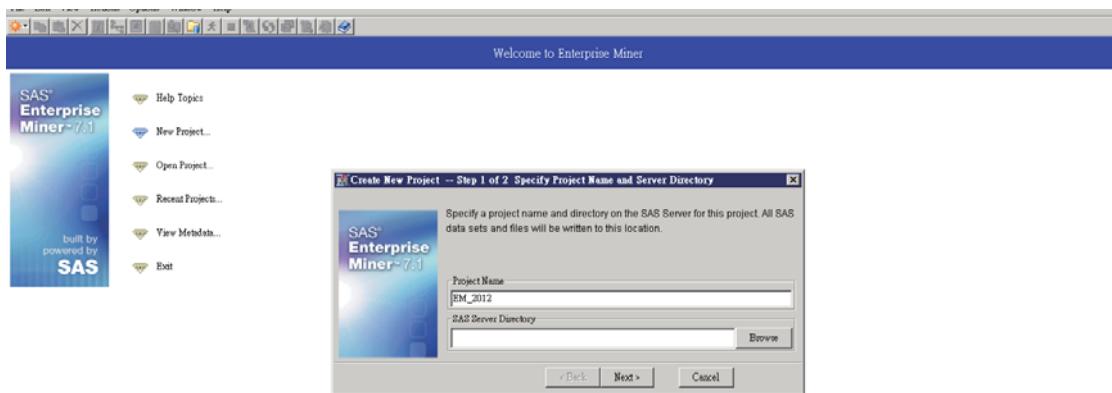
## A2: 附錄二 SAS Enterprise Miner 操作流程 Step by Step 畫面

A2-1 模組建置前置作業 .....	A2-1
1.1 新建專案(New Project).....	A2-1
1.2 新建類別庫(New Library) .....	A2-2
1.3 新建資料來源(Create Data Source).....	A2-4
1.4 新建圖表(New Diagram) .....	A2-10
A2-2 訓練模組建置 .....	A2-11
2.1 Replacement 節點.....	A2-12
2.2 資料分割(Data Partition).....	A2-13
2.3 決策樹(Decision Tree).....	A2-14
2.3.1 自動建置樹.....	A2-14
2.3.2 手動建置樹.....	A2-16
2.4 Impute 節點 .....	A2-20
2.5 迴歸(Regression)模型 .....	A2-21
2.6 類神經網路(Neural Network)模型 .....	A2-23
2.7 SVM 模型 .....	A2-24
2.8 模型比較(Model Comparison).....	A2-25
A2-3 評分 .....	A2-26
3.1 評分步驟.....	A2-26
3.2 評分結果.....	A2-28

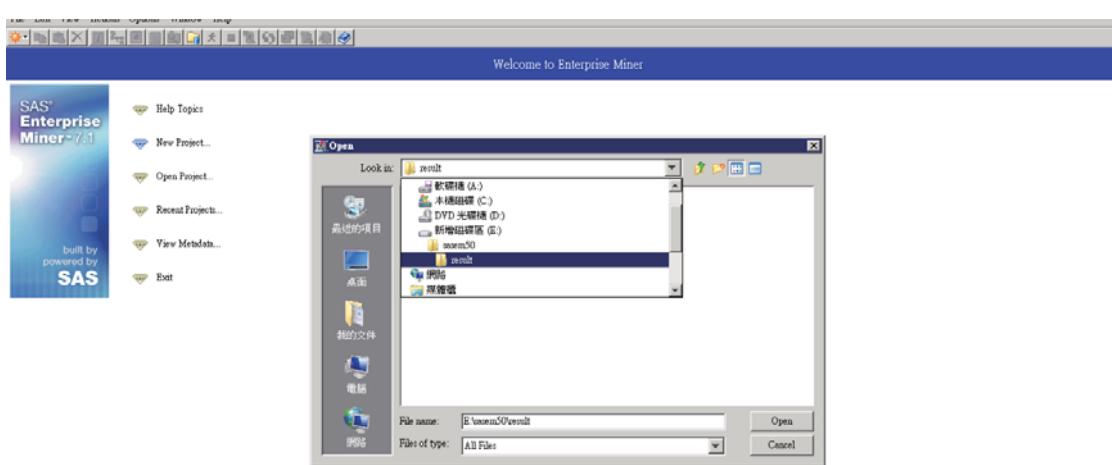
## A2-1 模組建置前置作業

### 1.1 新建專案(New Project)

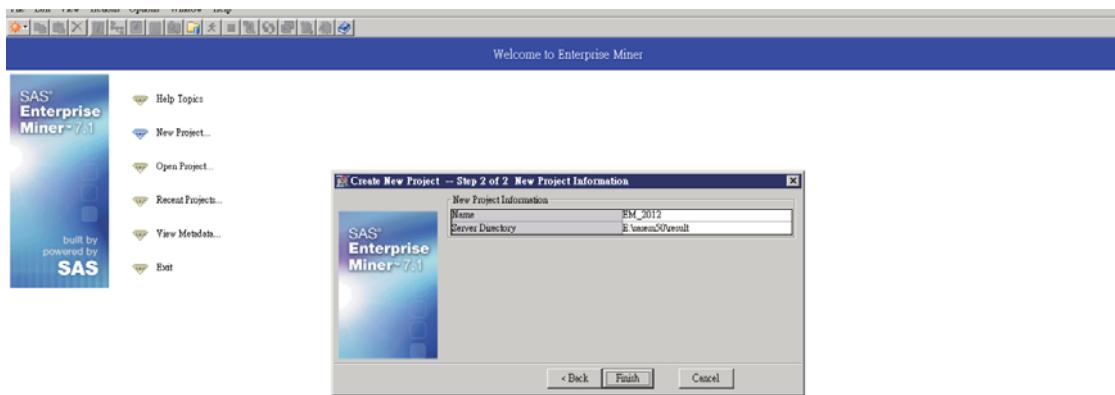
目的：模型建置之所有資訊將都被存在一個專案底下，因此首先須建置一個新專案。



Step1. 建立 New Project(新專案)，本專案名稱為 EM\_2012



Step2. 將本專案建置於路徑 E:\sasem50\result 下



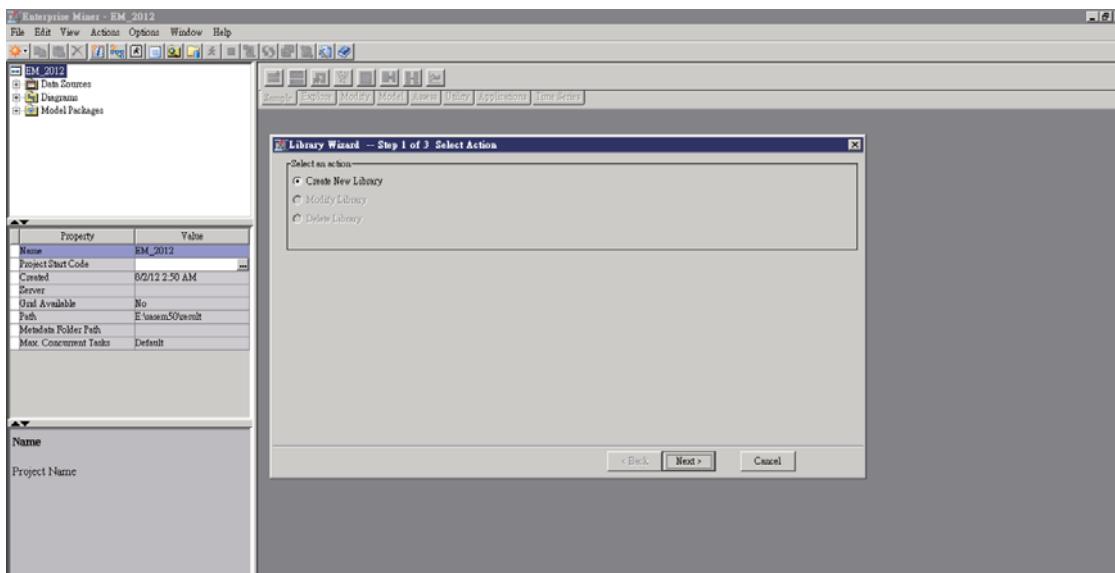
Step3.此為專案建置完成之資訊畫面，可確認名稱與所建之目錄。

## 1.2 新建類別庫(New Library)

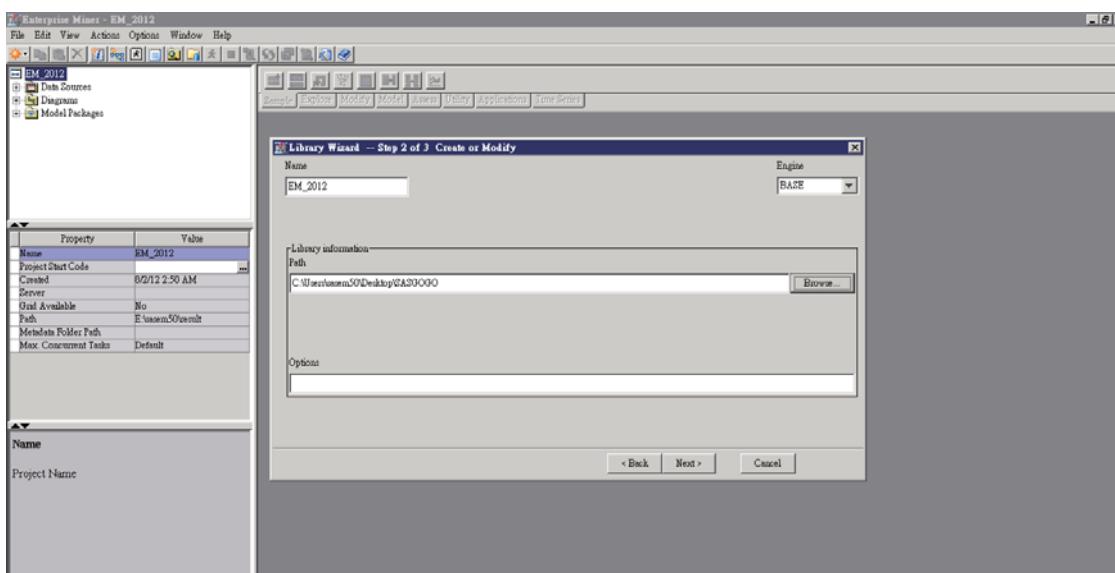
目的：類別庫建置目的主要用以結合專案所需之資料集以作為分析資料的基礎。



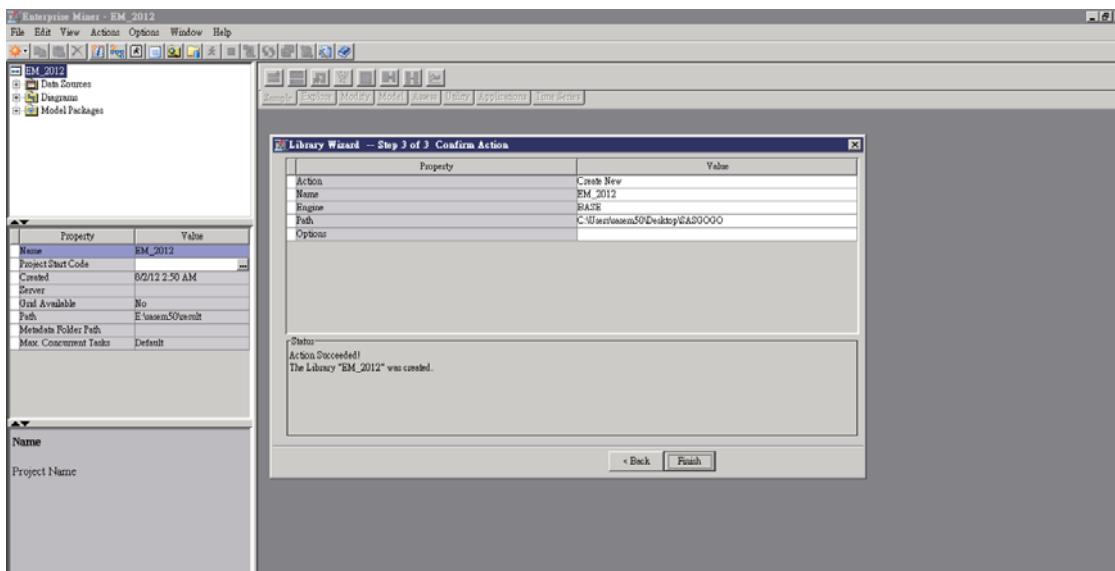
Step1.為了匯入 EG 所建之資料表，點選功能表上 File→New→Library 新建一個 Library(類別庫)



Step2. 選擇 Create New Library。



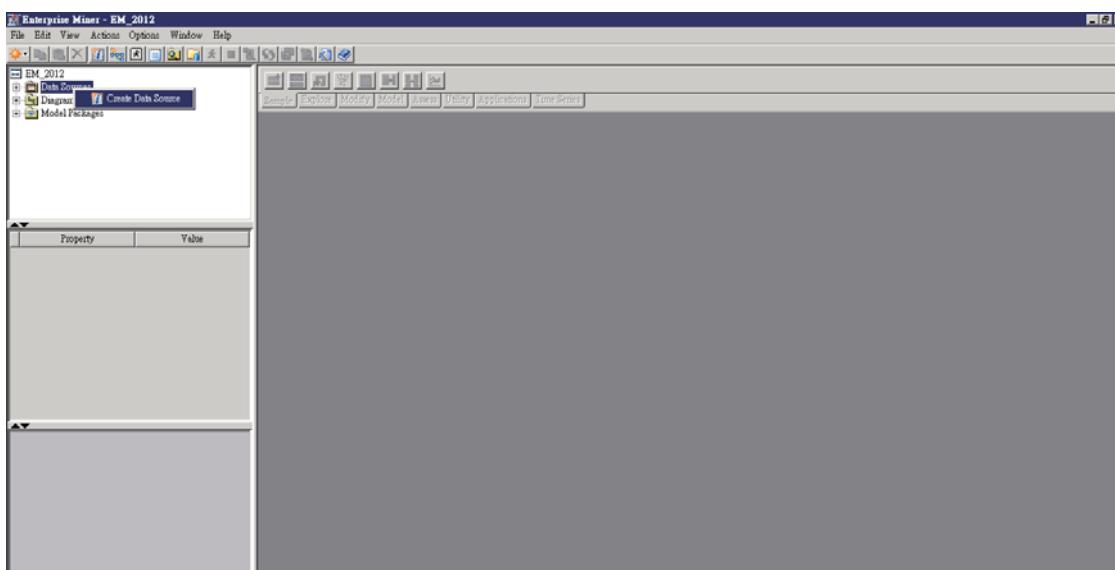
Step3. 輸入類別庫名稱，本類別庫名稱為 EM\_2012，在 Path 輸入框須輸入將匯入的資料表存在路徑。



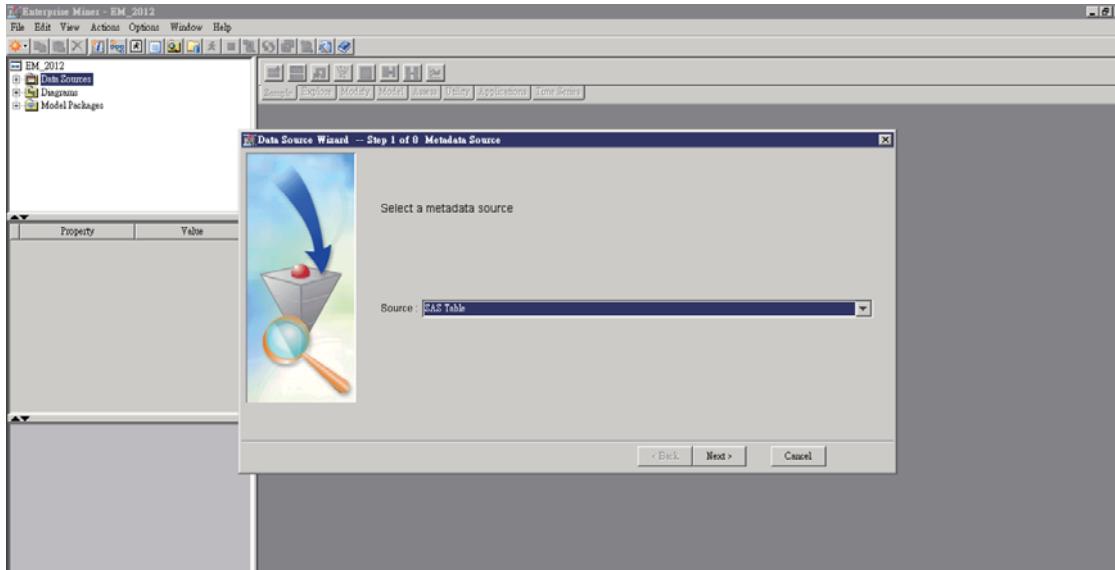
Step4. 最後確認所建類別庫名稱、路徑等資訊。

### 1.3 新建資料來源(Create Data Source)

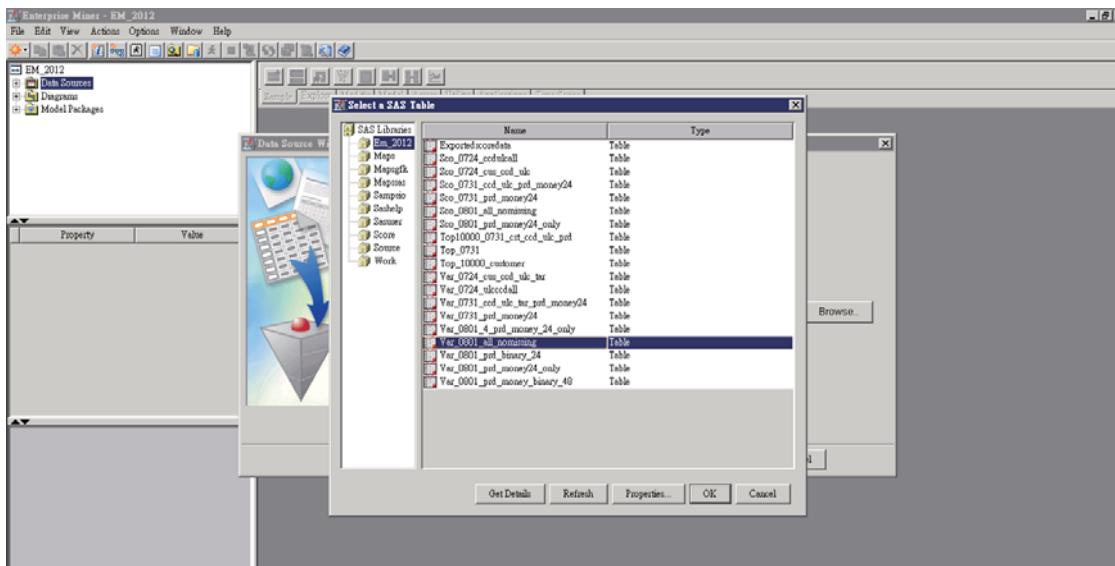
目的：匯入做預測模型所使用之資料集。



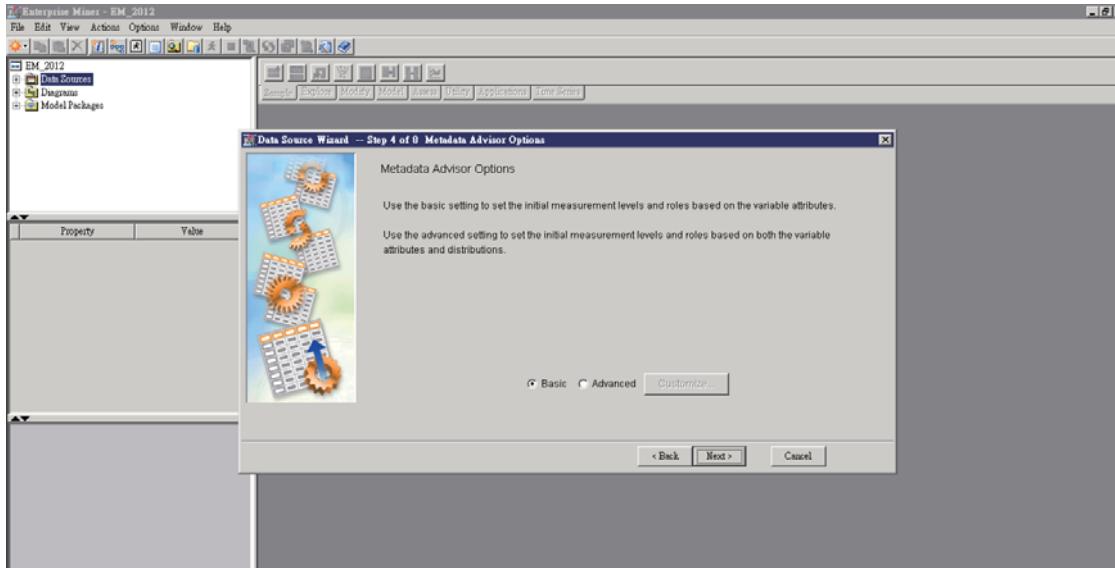
Step1.在左邊目錄的 Data Source 按右鍵→Create Data Source



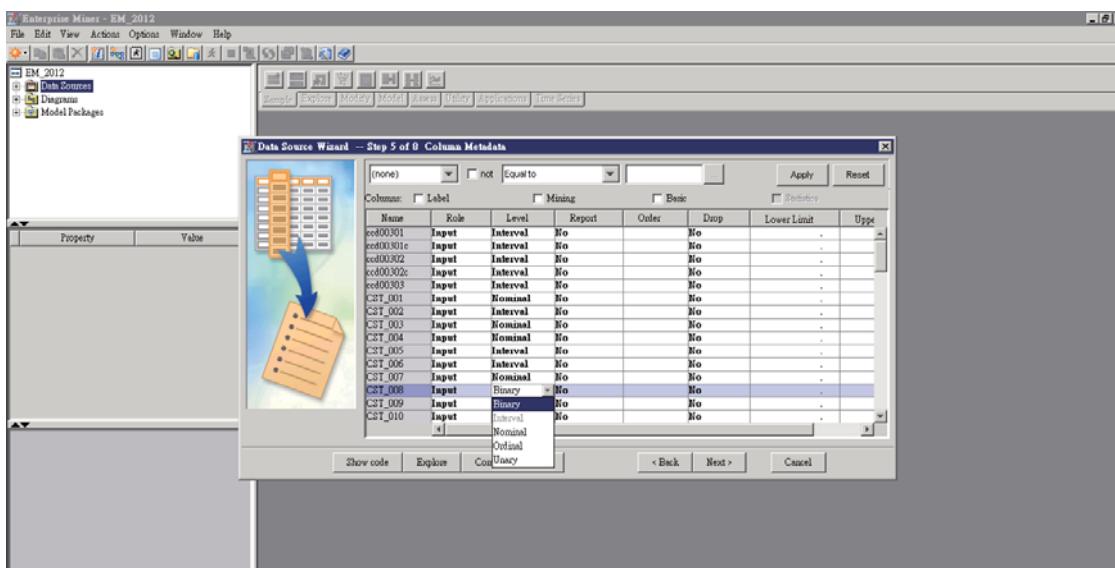
Step2.: 選擇中繼資料來源，來源設為 SAS Table 即可進行下一步。



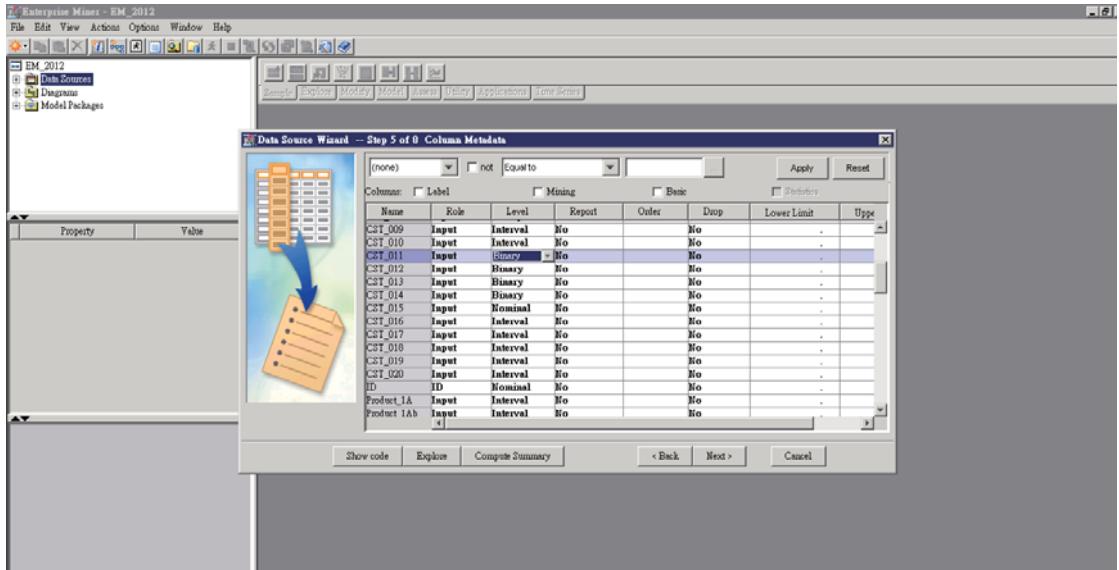
Step3. 於畫面左邊類別庫(Library)目錄中選取先前所建置之 Em\_2012 類別庫，將 VAR\_0801\_ALL\_NOMissing 資料表匯入，此資料表為我們將建置預測模型之重要資料來源。



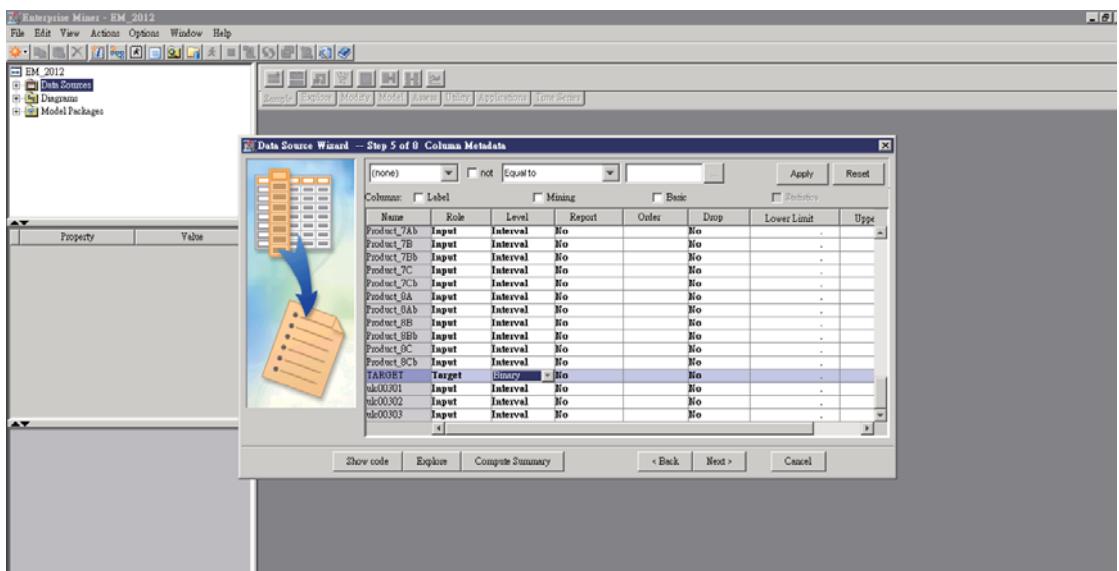
Step4. Metadata Advisor Options 有 Basic 與 Advanced 選擇，由於本專案只需更改基本的變數屬性，因此選擇 Basic 即可進行下一步。



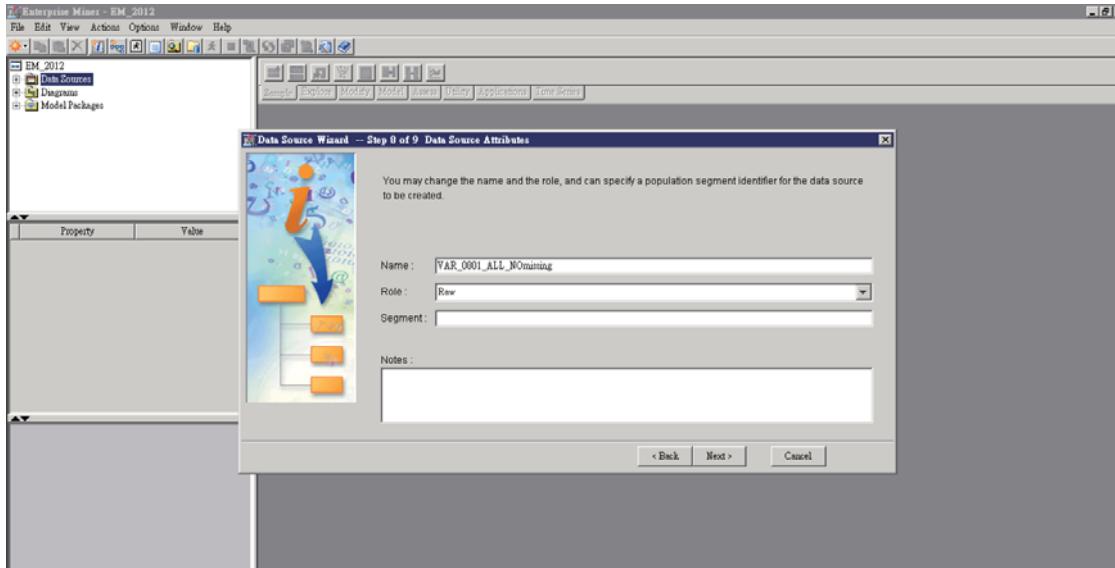
Step5. 進行個別欄位資料變數的層級(level)的調整，經觀察後應調整成正確的層級。以此圖為例，CST\_008 欄位描述為是否擁有附卡，因此須將欄位調整成 Binary(二元值)。



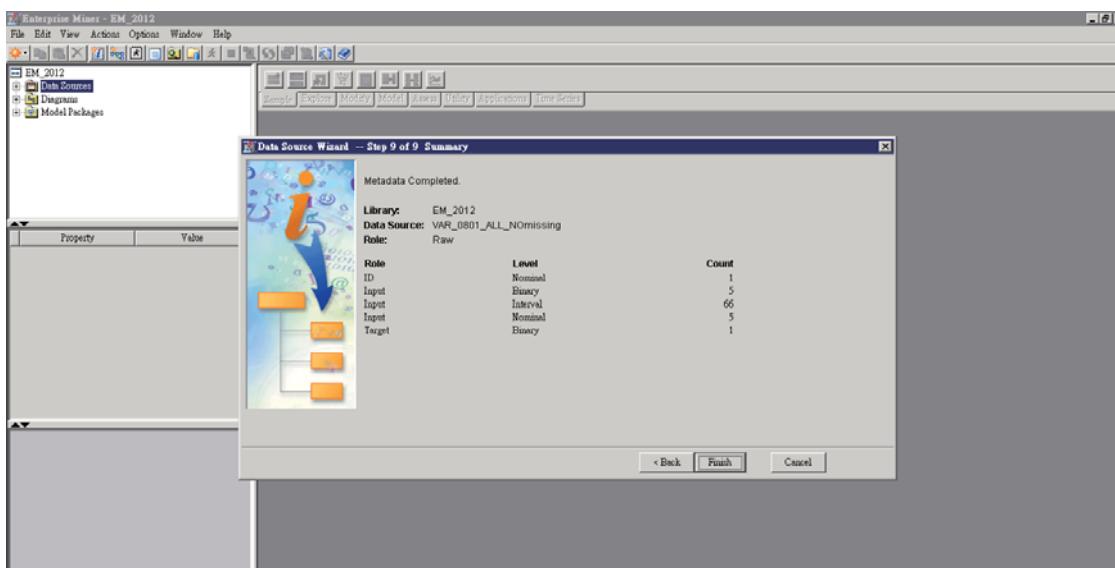
Step6. 進行個別欄位資料變數的層級(level)的調整，經觀察後應調整成正確的層級。以此圖為例，CST\_011 欄位描述為分期消費金註記(是否有分期消費)，因此須將欄位調整成 Binary(二元值)。



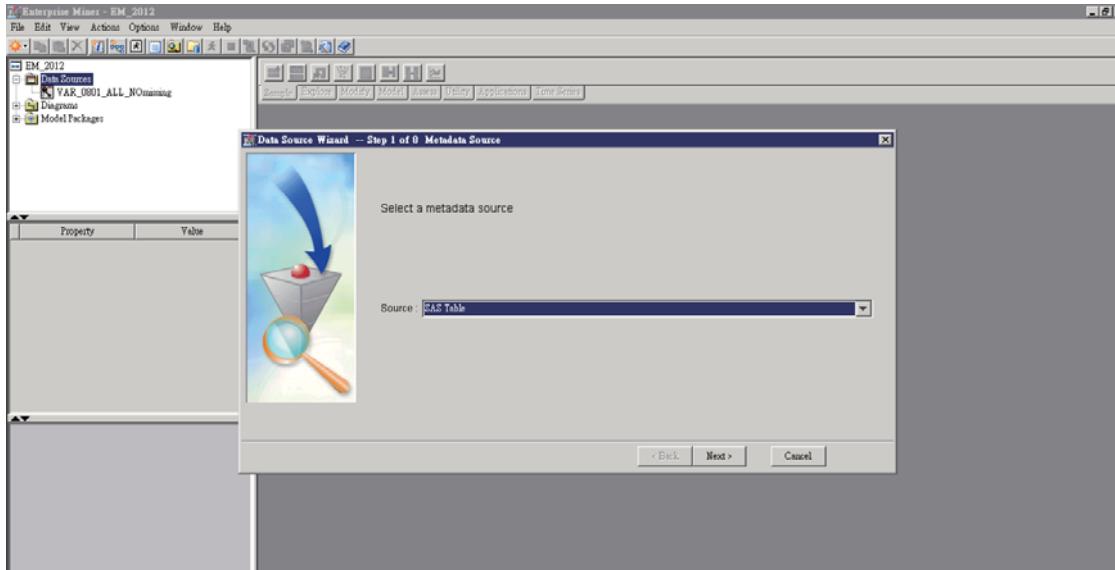
Step7. 進行個別欄位資料變數的層級(level)的調整，經觀察後應調整成正確的層級。以此圖為例，Target 欄位描述為預測目標變數-行銷活動實際成交客戶(有無成交)，因此須將欄位調整成 Binary(二元值)，最後將所有欄位變數確認完即可進行下一步。



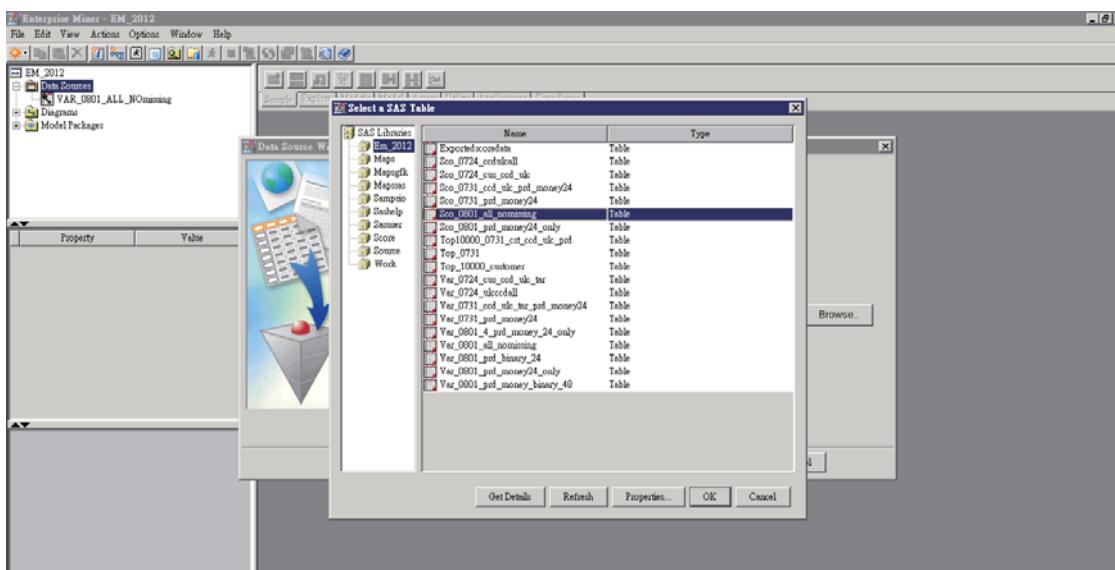
Step8. 由於本專案使用資料表為原始資料(Raw)的型態，故資料表角色選擇 Raw 即可進行下一步。



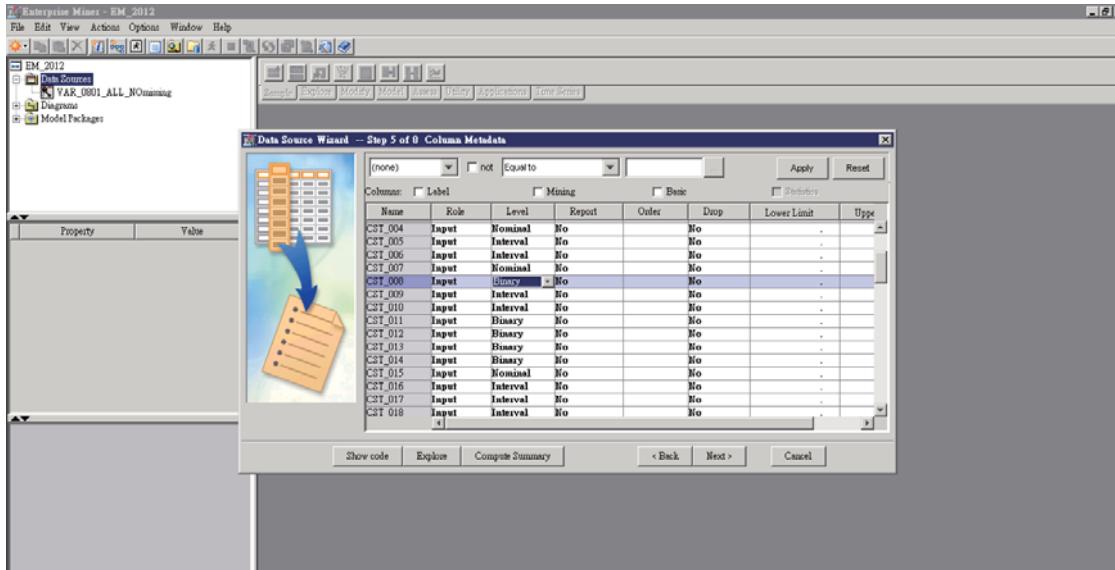
Step9. 資料來源建置完成之資訊，除了 Library、Data Source 與 Role 的資訊確認，不同層級的變數調整結果總數亦可由此圖所見。



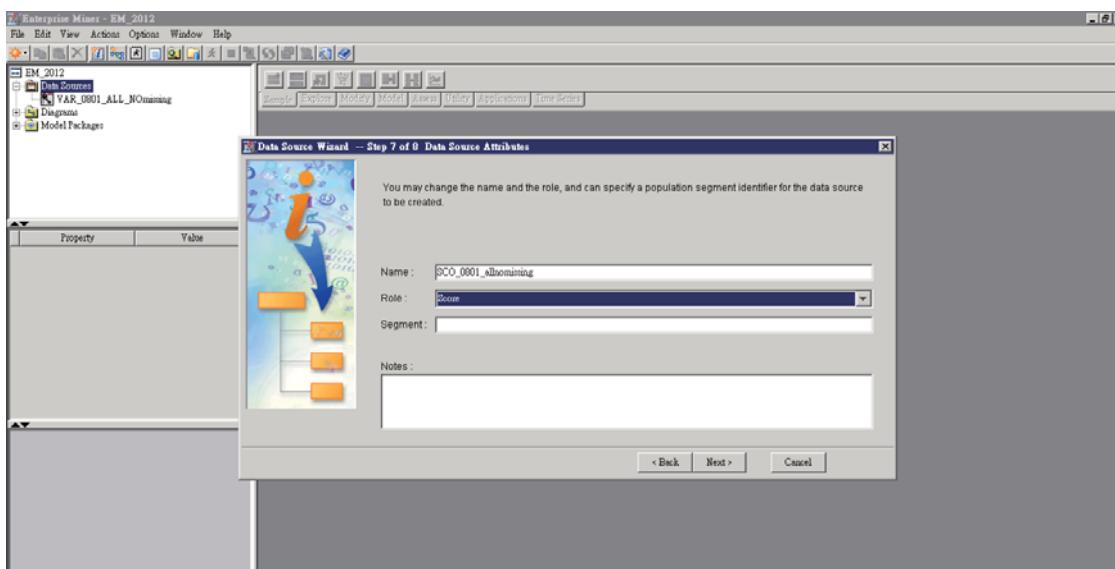
Step10. 由於本專案目的為將過去的資料做未來資料的預測，因此須將評分資料(即將要預測的資料)匯入，資料來源一樣設定 SAS Table 即可進行下一步。



Step11. 於畫面類別庫(Library)目錄中選取先前所建置之 Em\_2012 類別庫，將 Sco\_0801\_all\_nomissing 資料表匯入，此資料表為欲預測之資料表。



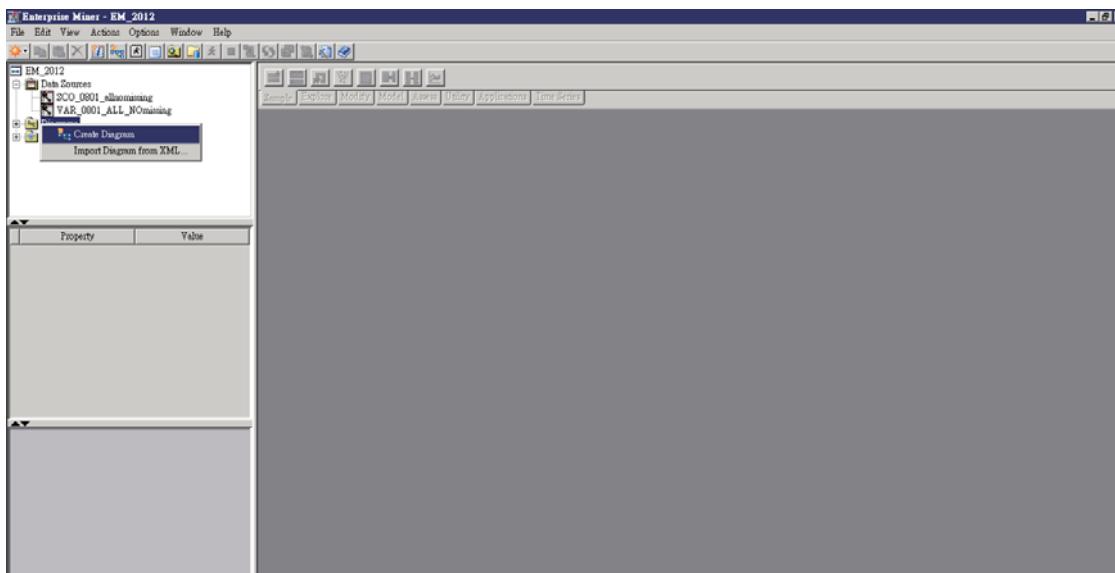
Step12. 同訓練資料，評分資料亦須進行個別欄位資料變數的層級(level)的調整，經觀察後應調整成正確的層級。以此圖為例，CST\_008 欄位描述為是否擁有附卡，因此須將欄位調整成 Binary(二元值)，最後把全部資料表欄位檢查完畢之後即可進行下一步。



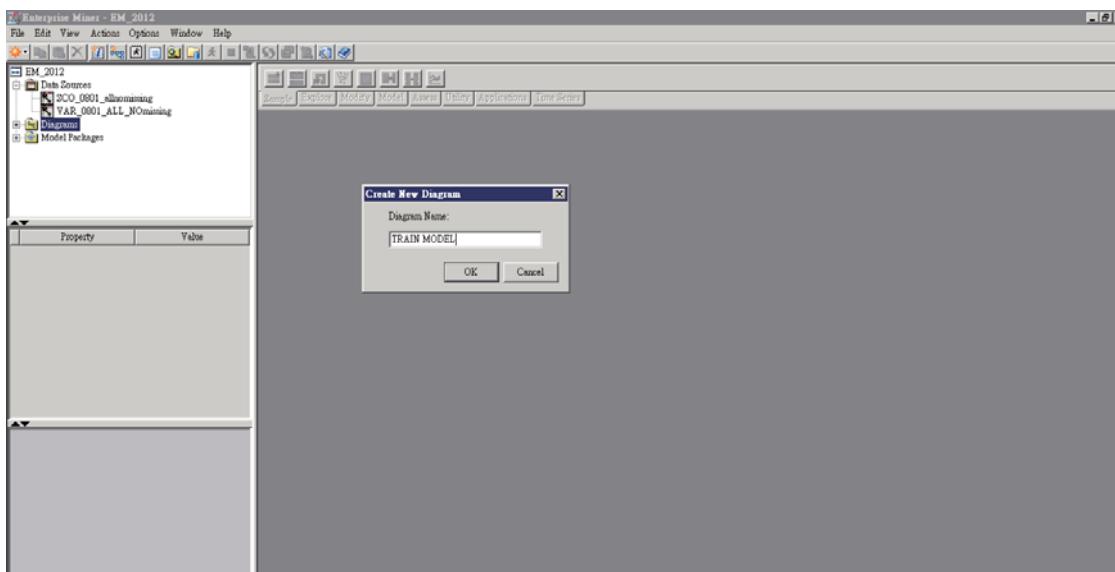
Step13. 由於本專案使用評分資料表亦為原始資料(Raw)的型態，故資料表角色選擇 Raw 即可進行下一步，最後亦確認完建置資訊即完成資料來源匯入之所有動作。

## 1.4 新建圖表(New Diagram)

目的：接下建置的模型皆須放置在圖表上。



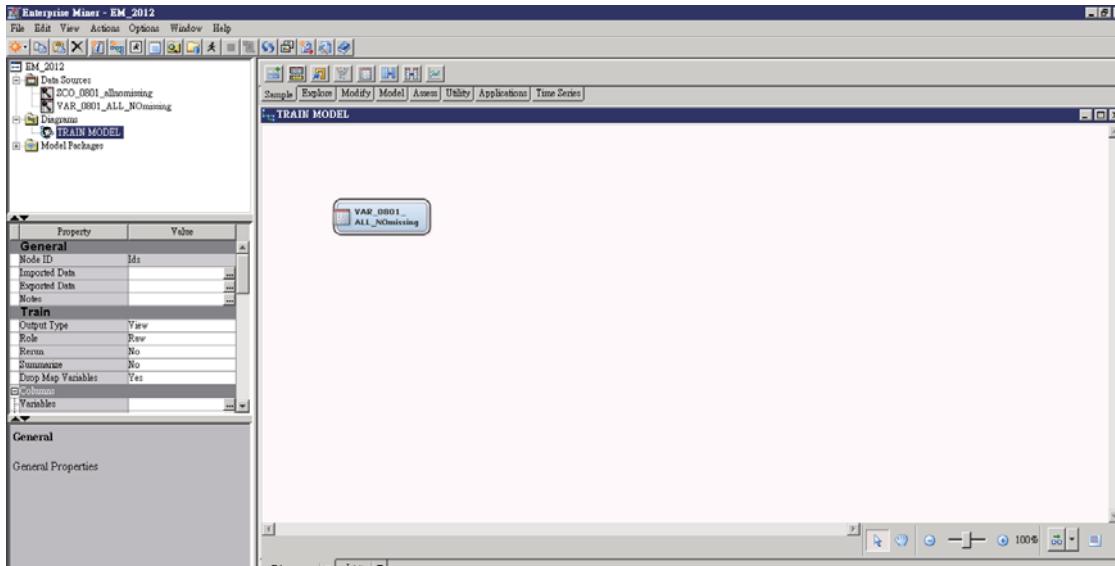
Step1 在左邊目錄視窗 Diagram 圖示按右鍵→Create Diagram(建立圖表)



Step2 在 Create New Diagram 視窗輸入框輸入欲建置的圖表名稱，本專案設定圖表名稱為 TRAIN MODEL。

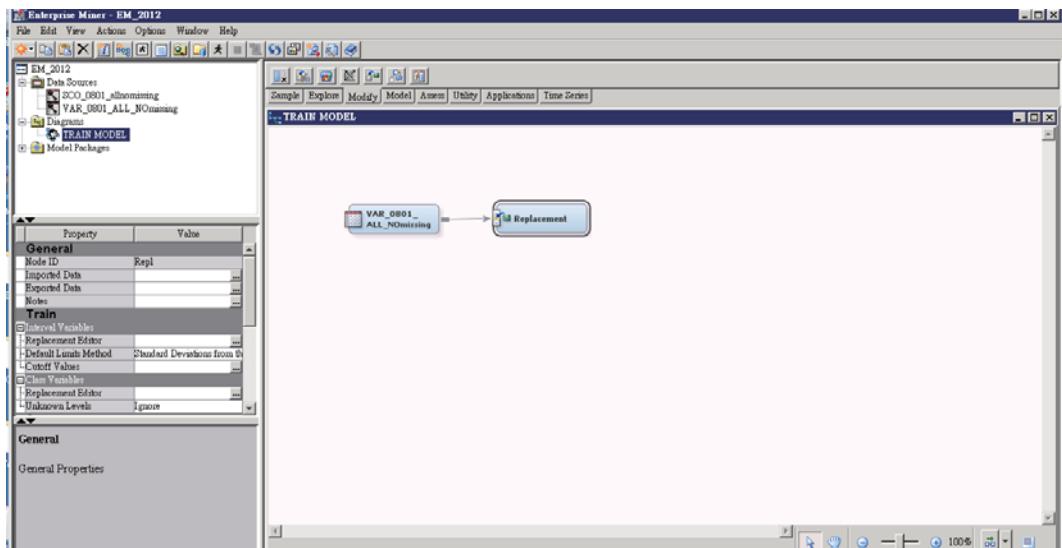
## A2- 2 訓練模組建置

目的：嘗試許多節點方法做成預測模型。

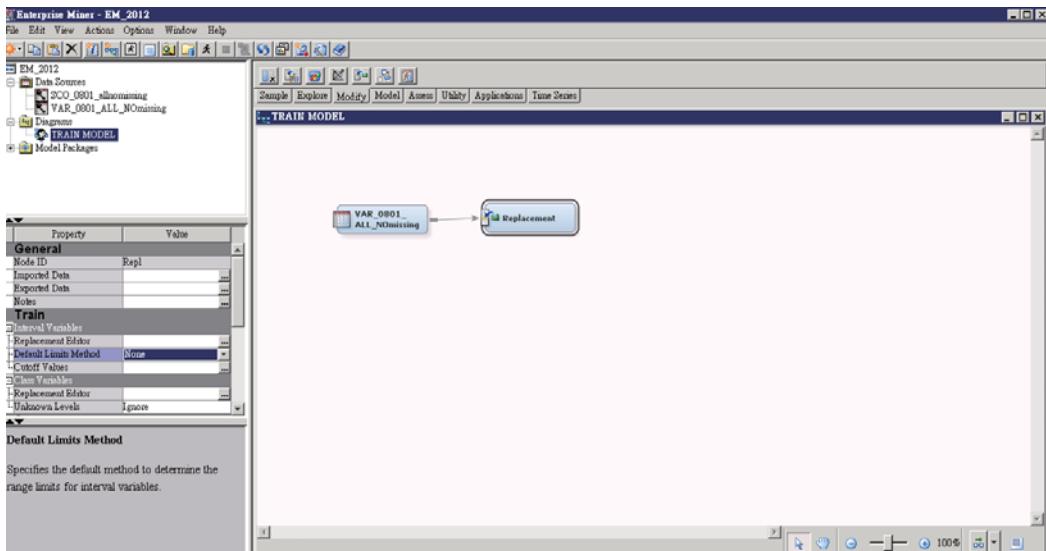


在開始建置模型前，由於是以資料表來源進而透過預測模型達到模型建置的目的，因此須將先左邊 Data Source 目錄所匯入的資料表 VAR\_0801\_ALL\_NOMissing 拖曳至在 TRAIN MODEL 圖表中。

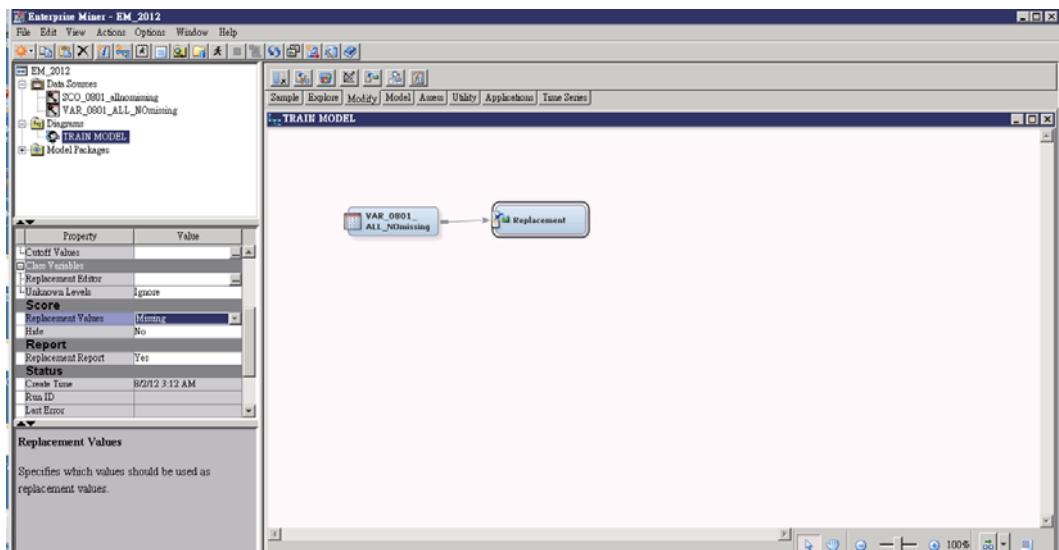
## 2.1 Replacement 節點



**Step1.** 接下來所使用之功能皆為 SEMMA 的節點，首先使用 Modify 的 Replacement 節點，將它拖曳於圖表中，目的將遺失值做取代的動作。



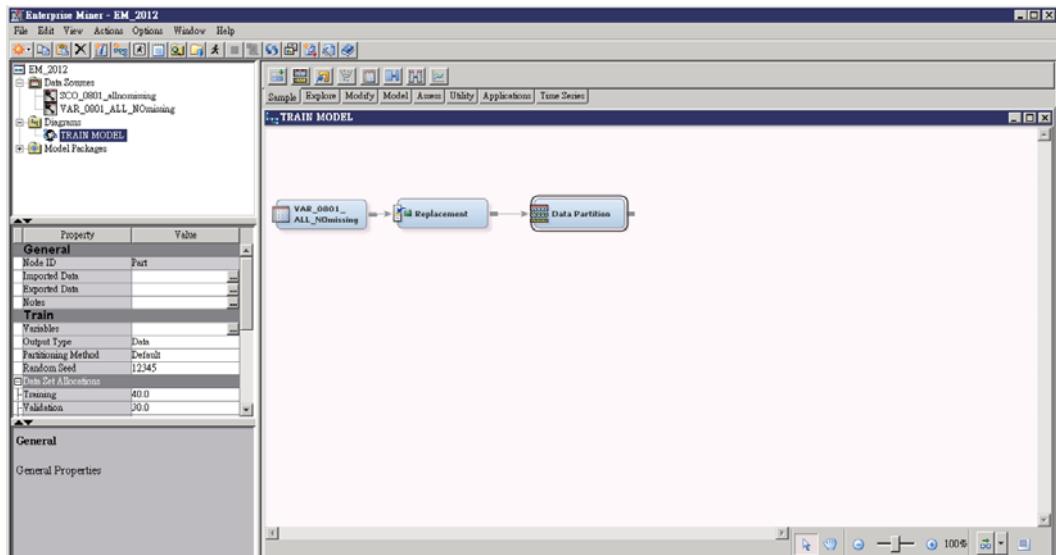
**Step2.** 在左邊屬性參數設定中，將 Train→Default Limits Method 設定為 None。



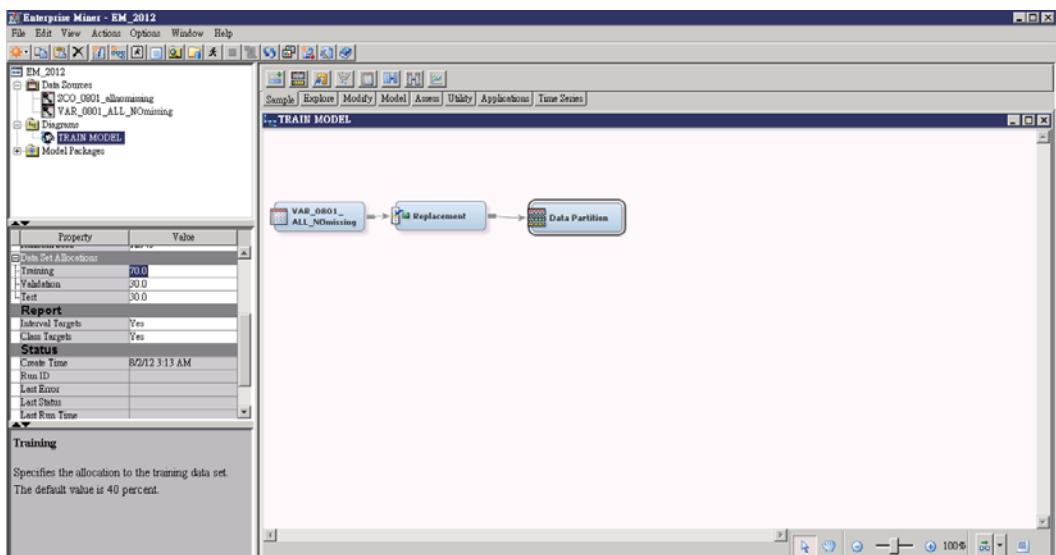
**Step3.** 在左邊屬性參數設定中，設定 Score→Replacement Value→Missing

## 2.2 資料分割(Data Partition)

目的：為建立一個預測模型，須將資料切割成訓練、驗證及測試的三個樣本，此部分本專案將資料切割成訓練及驗證兩個部分。



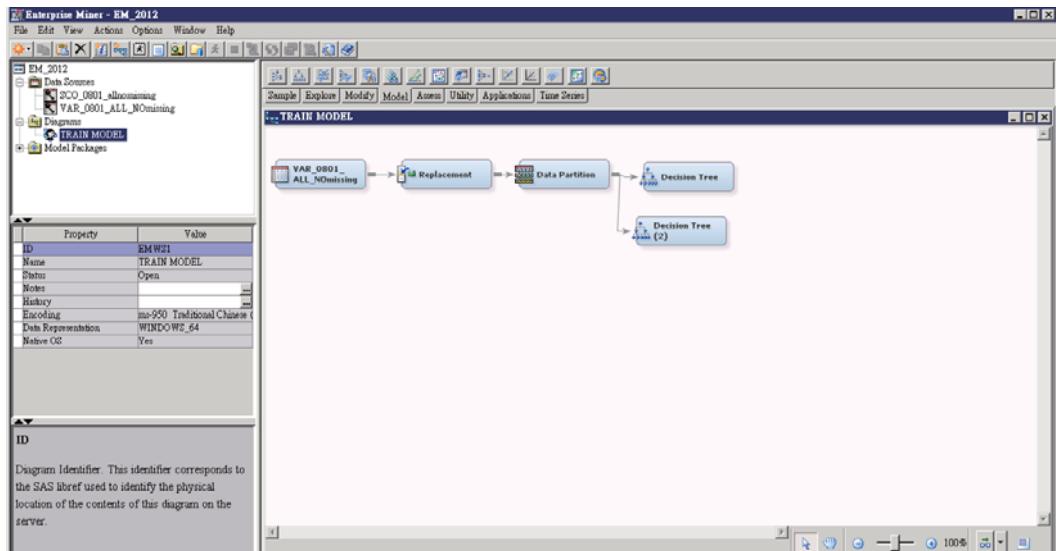
**Step1.** 將 Sample→Data Partition(資料分割)節點拖曳置圖表中，接下來將進行資料分割的動作。



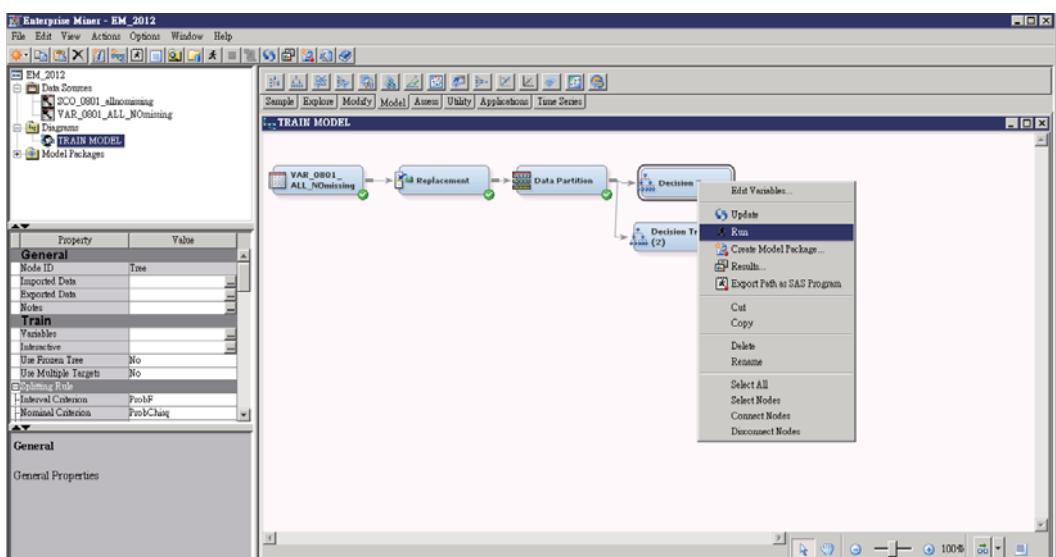
**Step2.** 在左邊屬性視窗的 Data Set Allocations 進行參數調整，本專案的資料配置將資料切割為訓練與驗證兩部分，而根據過往研究將比例調成 7:3 為最好的訓練方式，故將訓練與驗證分別調成 70%與 30%。

## 2.3 決策樹(Decision Tree)

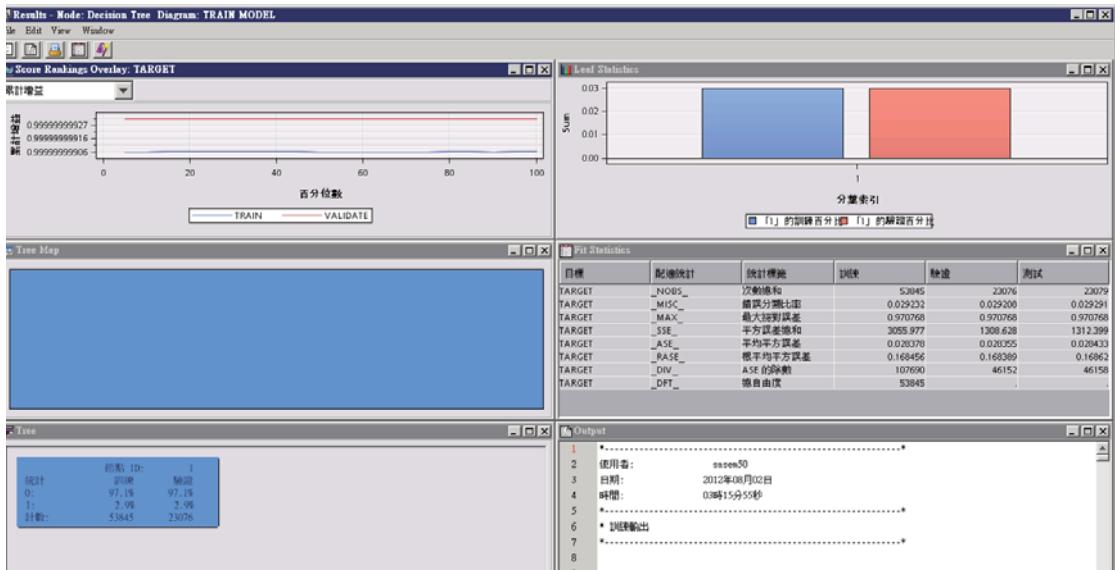
### 2.3.1 自動建置樹



**Step1.** 將資料分割後，將 Model→Decision Tree(決策樹)節點拖曳置圖表中以建置模型，本專案將使用兩個決策樹建立預測模型，Decision Tree 為自動建立的樹，Decision Tree(2)為手動建立的樹。

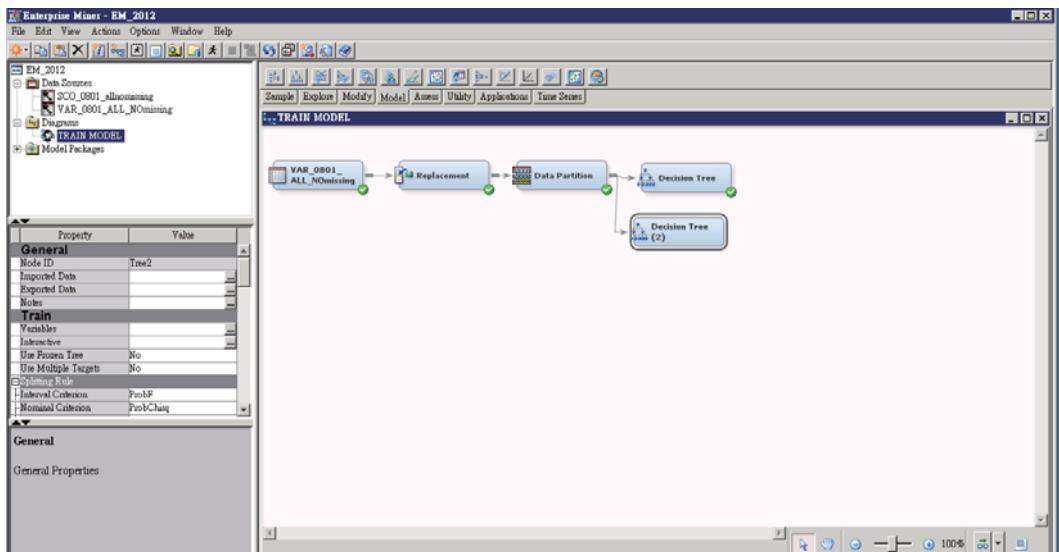


**Step2.** 在 Decision Tree 點選右鍵→Run 可直接執行自動建立的樹。

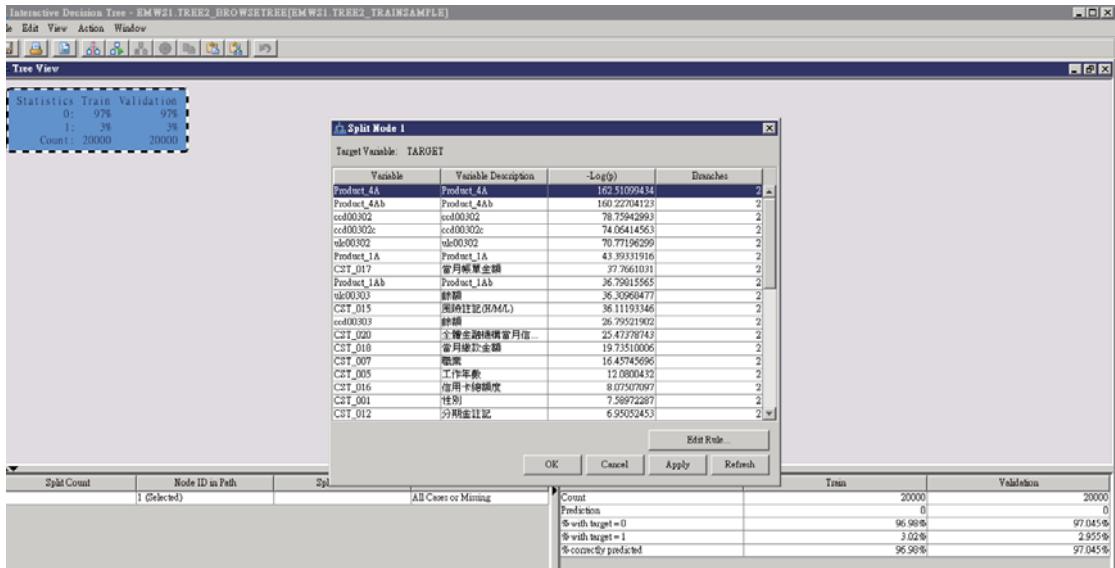


**Step3.** 此為 Decision Tree 自動建立的樹結果，輸出的結果系統並未作分割的動作，因此可得知自建樹是不建議作分割的。

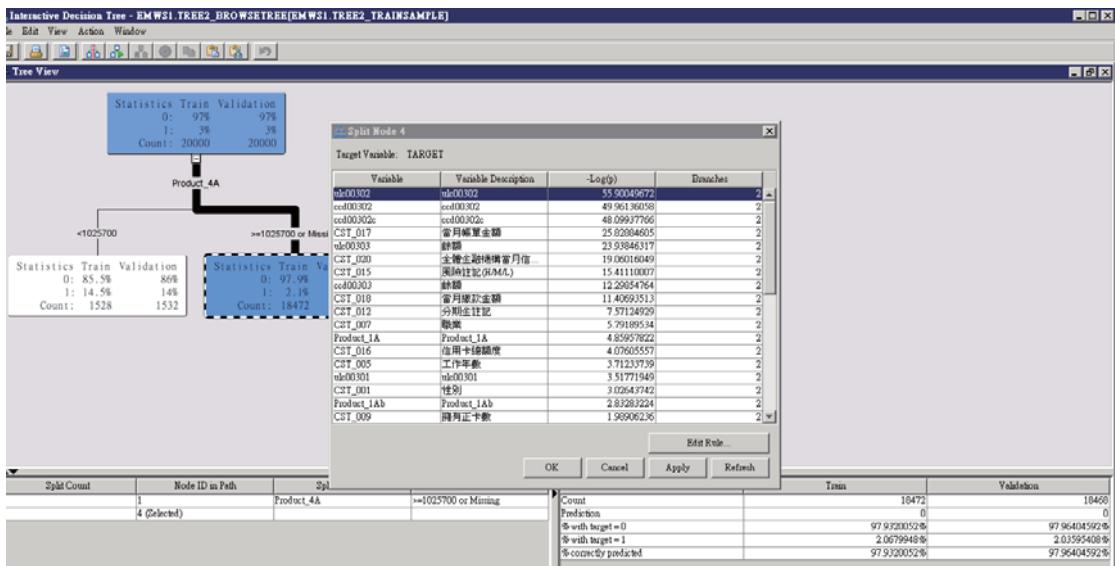
### 2.3.2 手動建置樹



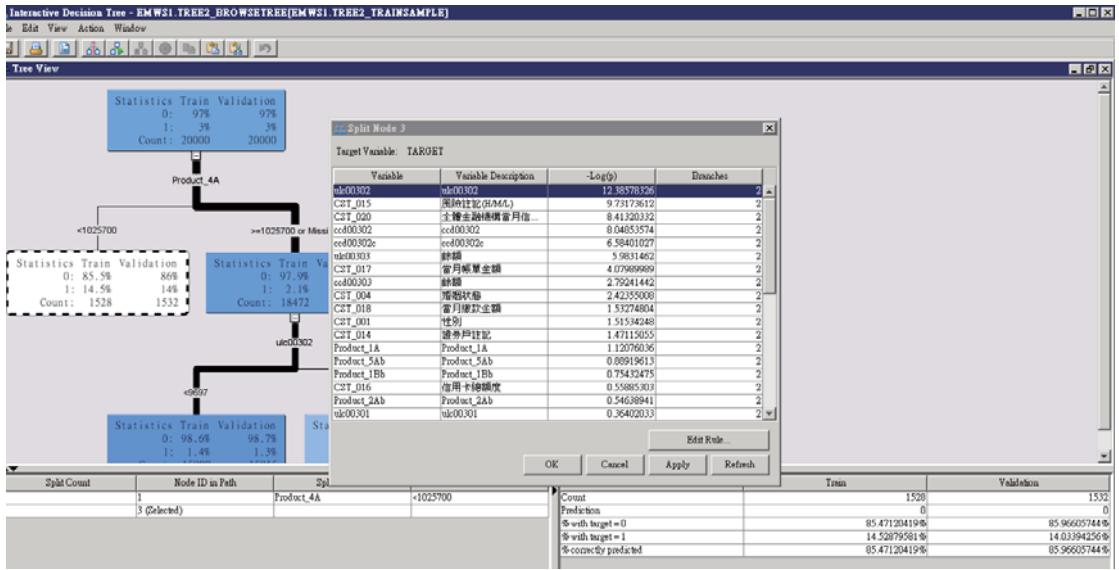
**Step1.** 接著進行 Decision Tree(2)自建樹的部分，首先須在左邊屬性區 Train→Interactive 即可進行手動建置。



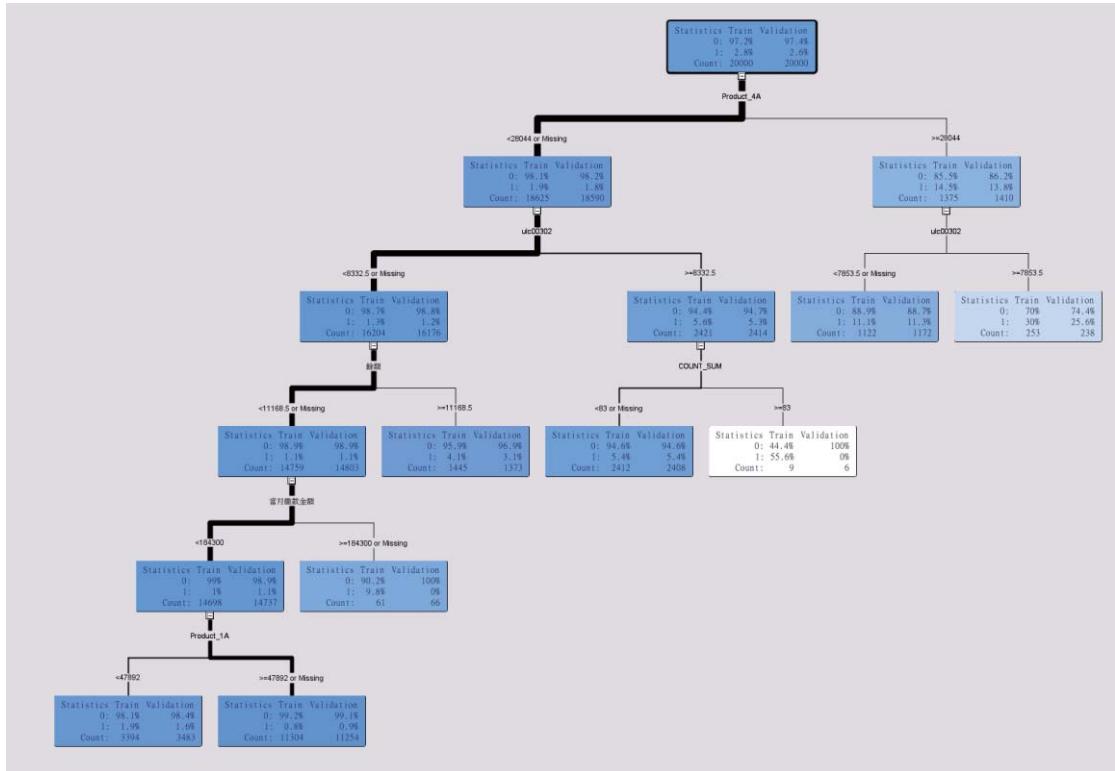
Step2. 手動建置樹我們針對 $-\log(p)$ 的大小做分割， $-\log(p)$ 愈大愈具有影響力，分割依據為針對資料筆數多者與 $-\log(p)$ 值較大的往下分割，本專案於 $-\log(p)$ 小於5時就不再做分割。此圖為分割的第一個變數欄位 Product\_4A 信貸(正常)，首先點選此變數按右鍵進行分割。



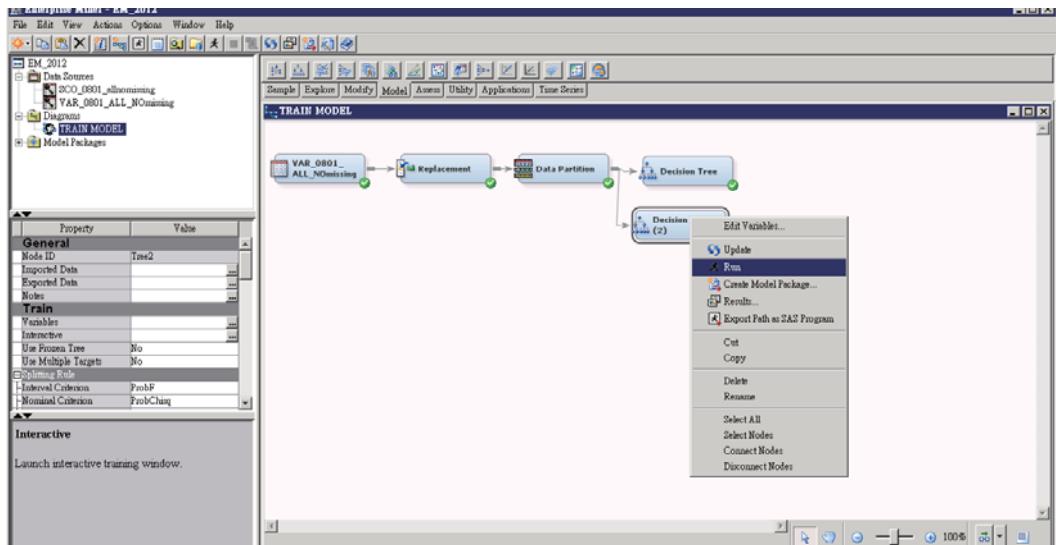
Step3. 此為分割的第二個變數欄位—ulc00302(行外全體金融機構預借現金金額)，同上可點選此變數按右鍵進行分割。



Step4. 此為分割的第三個變數欄位—ulc00302(行外全體金融機構預借現金金額)分割結果，與上步驟選擇變數雖相同，但其分割的對象不同，同前步驟一樣點選右鍵進行分割。依此類推本專案的自動建置樹總共以五個變數作分割樹的依據分別為：Product\_4A 信貸(正常)、ulc00302(行外全體金融機構預借現金金額)、餘額、當月應繳金額與 Product\_1A 存款(正常)。



Step5. 此圖為手動建置樹以 $-\log(p)$ 分割之最後圖形結果。

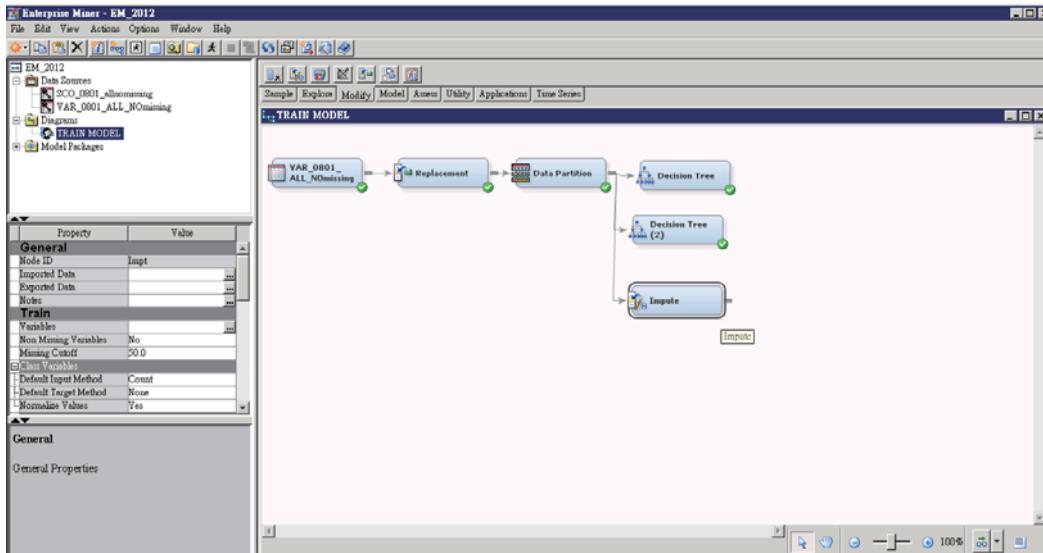


Step6. 在 Decision Tree(2)點選右鍵→Run 即可看輸出結果。

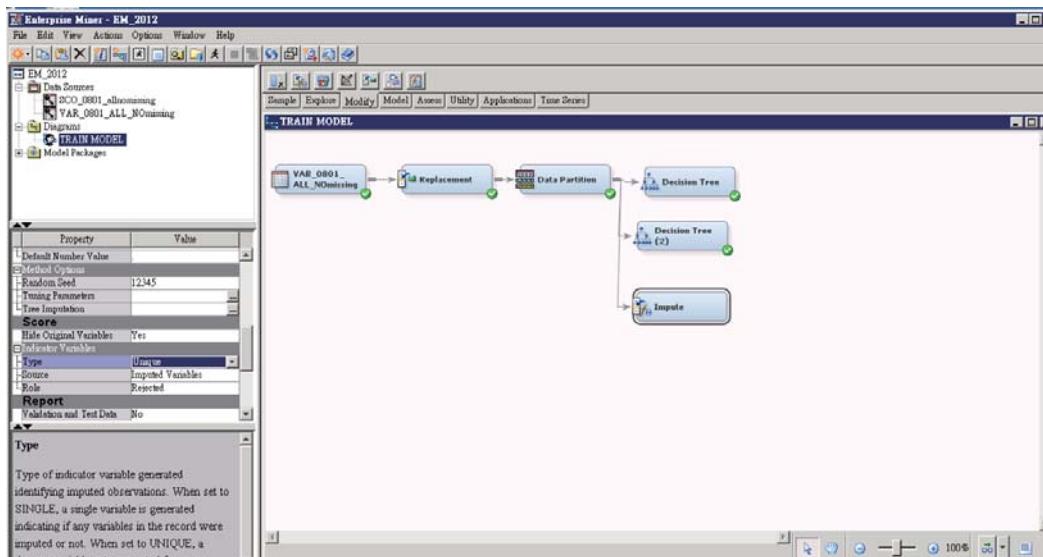


Step7. 此圖為點擊輸出結果之增益表。

## 2.4 Impute 節點

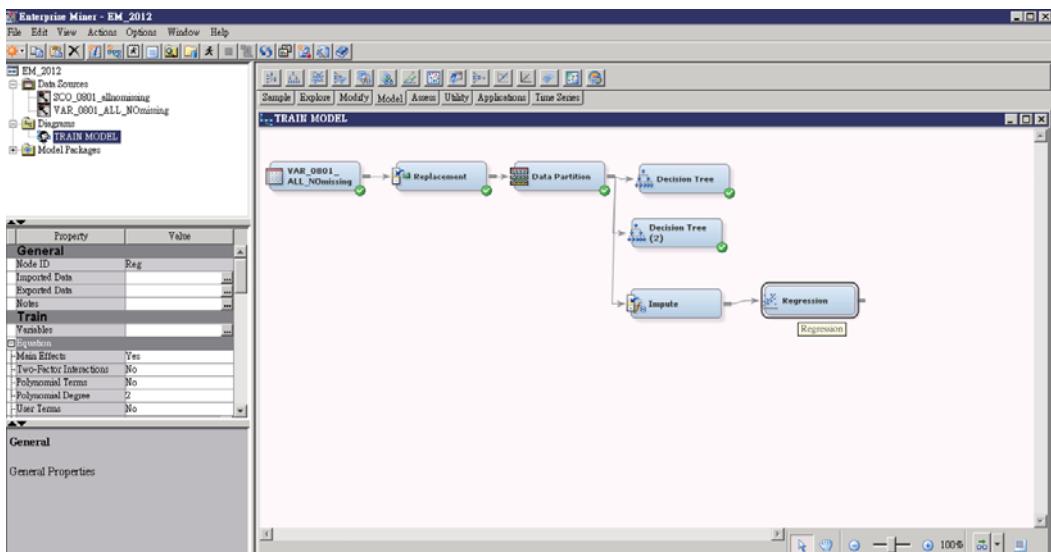


**Step1.** 由於接著將建立回歸(Regression)節點做預測模型，但當資料集有遺失值時，會對模型有影響，因此將 Modify→Impute 拖曳至圖表中，使用 Impute 節點先做遺失值取代的動作，

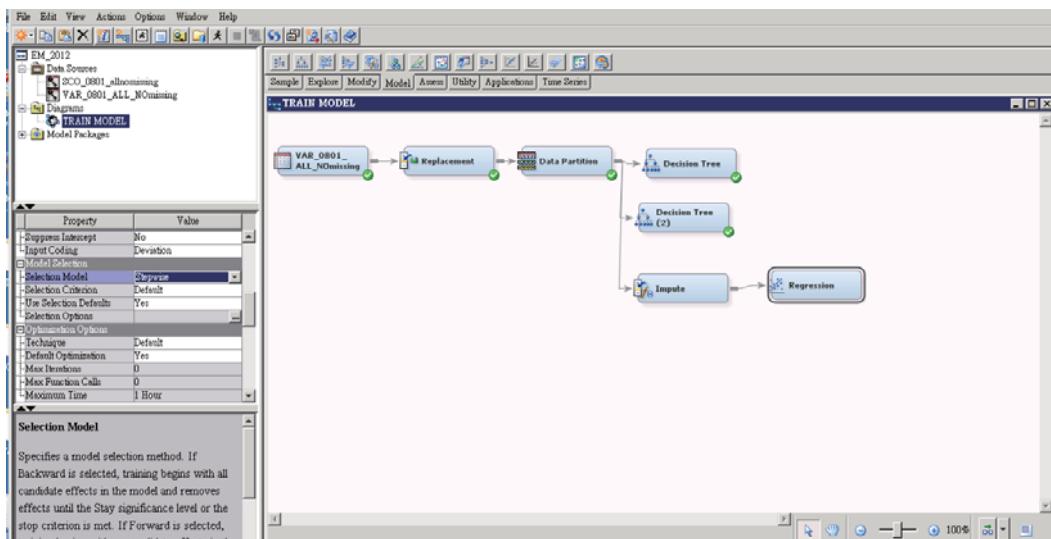


**Step2.** 在左邊屬性參數視窗做調整，Indicator Variable→Type→Unique。

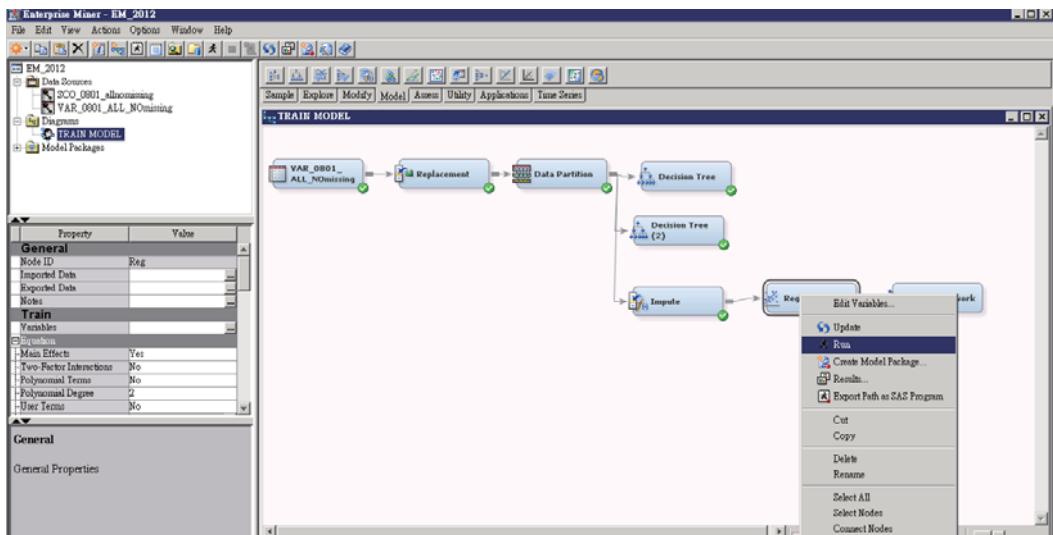
## 2.5 回歸(Regression)模型



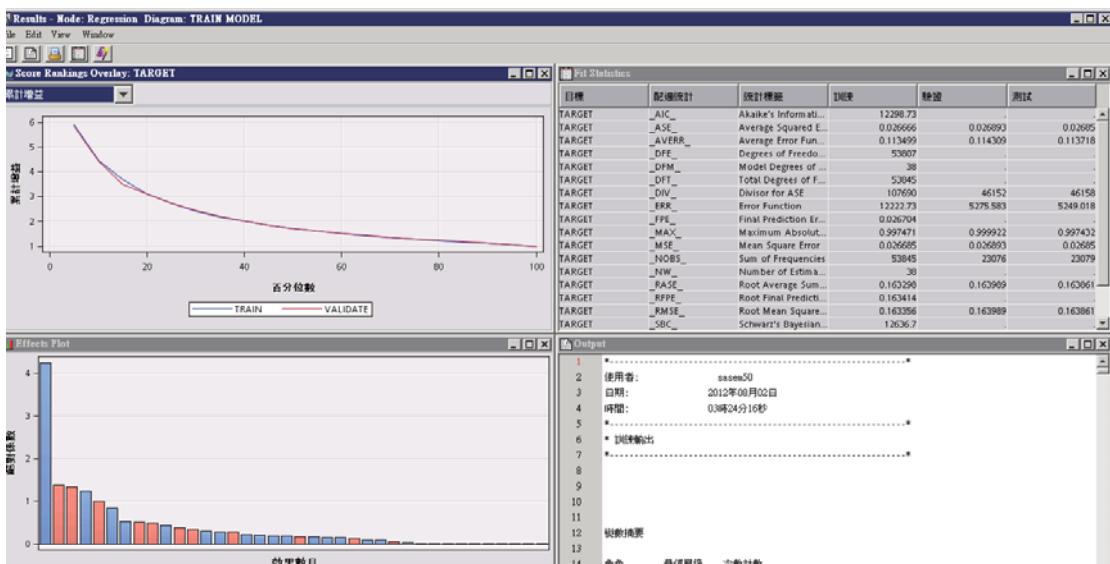
**Step1.** 經過 Impute 取代遺失值後，將 Model→迴歸(Regression)節點拖曳至圖表中以做預測模型。



**Step2.** 在左邊屬性參數視窗中的 Model Selection→Selection Model 將參數設定為 stepwise 方法，由於此方法綜合 forward 和 backward 方法且預測能力最好因此做為專案回歸模型的主要方法。

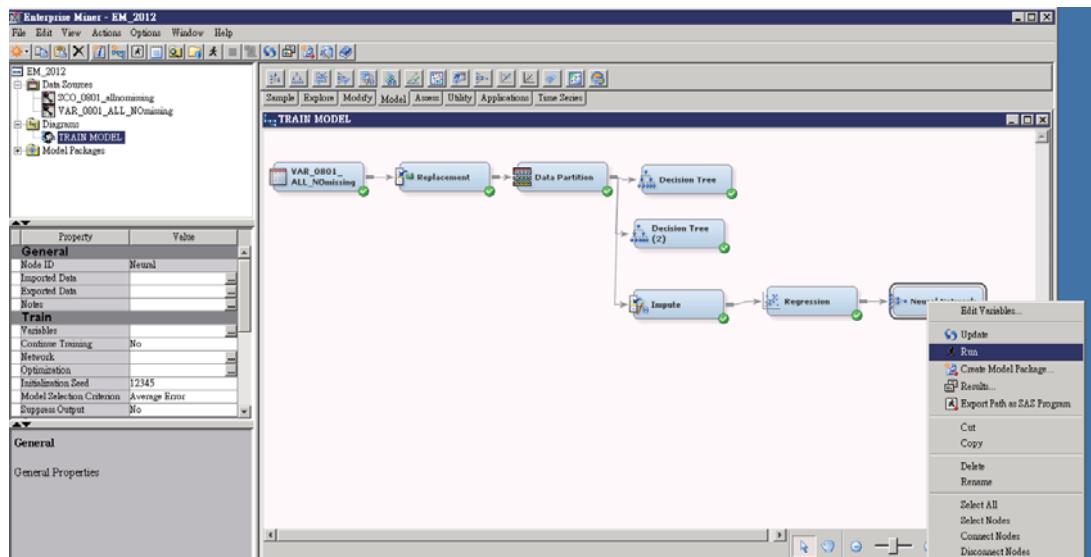


Step3. 在迴歸(Regression)節點按右鍵→Run，開始執行迴歸(Regression)結果。

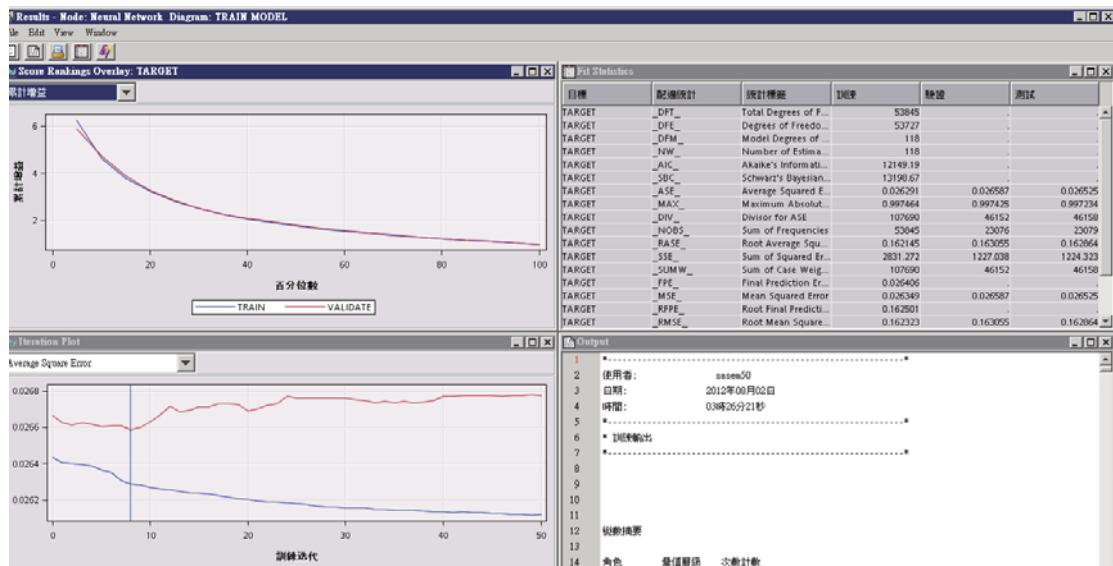


Step4. 此為迴歸(Regression)輸出之結果。

## 2.6 類神經網路(Neural Network)模型

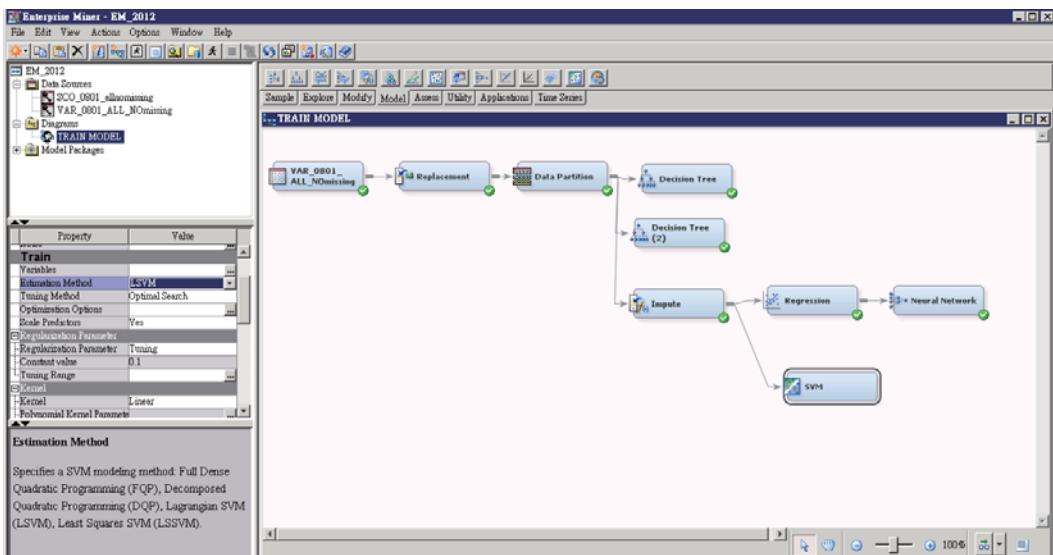


**Step1.** 選擇 Model→Neural Network 類神經網路節點，將之拉近圖表中以建立預測模型，於左邊視窗調整參 Train→Model Selection Criterion 選擇 Average Error 平均誤差方法。確認方法後再以拖曳進圖表中的 Neural Network 節點按右鍵→Run 即可執行結果。

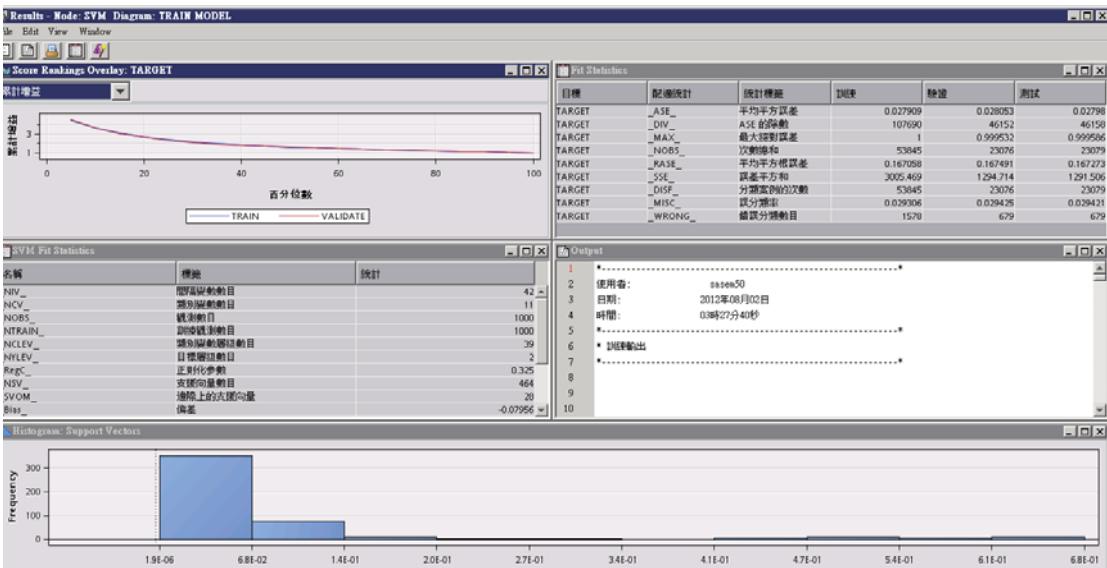


**Step2.** 此為類神經網路(Neural Network)節點輸出結果。

## 2.7 SVM 模型

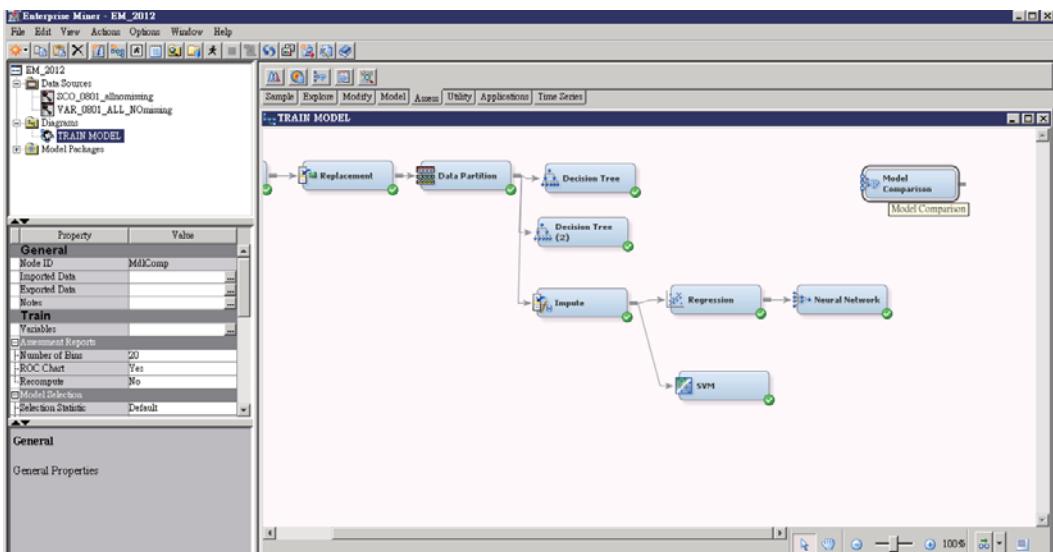


**Step1.** 本專案嘗試 SVM 模型，首先將 SEMMA 的 Model→SVM 拖曳至圖表中，本團隊在先前已測試了所有參數的組合，發現 LSVM 線性方法的效果最佳，因此在左邊參數調整視窗中，選擇 Train→Estimation→LSVM 做預測模型。確認方法後再以拖曳進圖表中的 SVM 節點按右鍵→Run 即可執行結果。

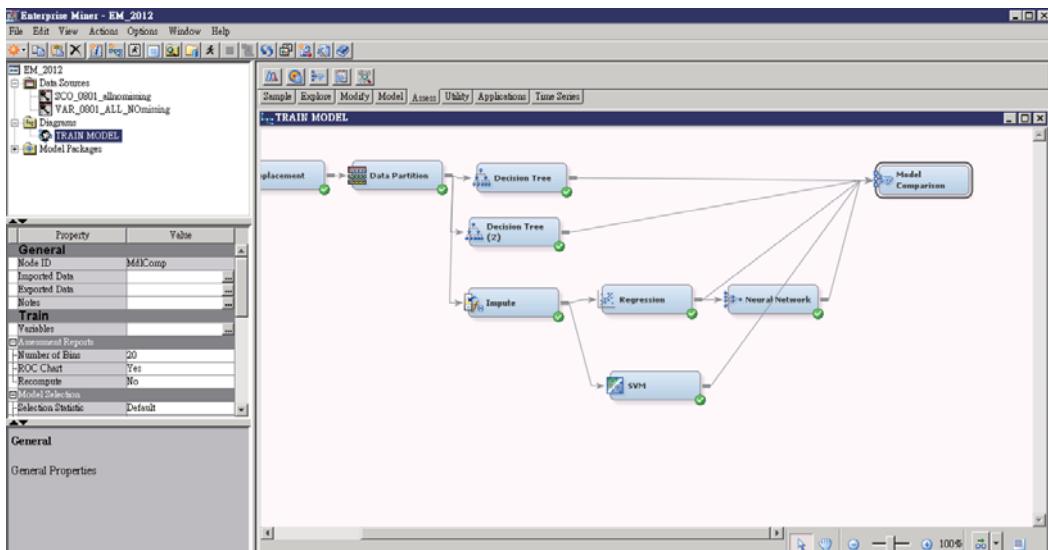


**Step2.** 此為 SVM 輸出之結果。

## 2.8 模型比較(Model Comparison)



**Step1.** 訓練完模型後，本專案利用 Model Comparison(模型比較)節點進而比較五個模型的好壞。將 Access→Model Comparison 拖曳至圖表中。



**Step2.** 將五的預測模型 Decision Tree、Decision(2)、Regression、Neural Network、SVM 箭頭拉至 Model Comparison(模型比較)節點。

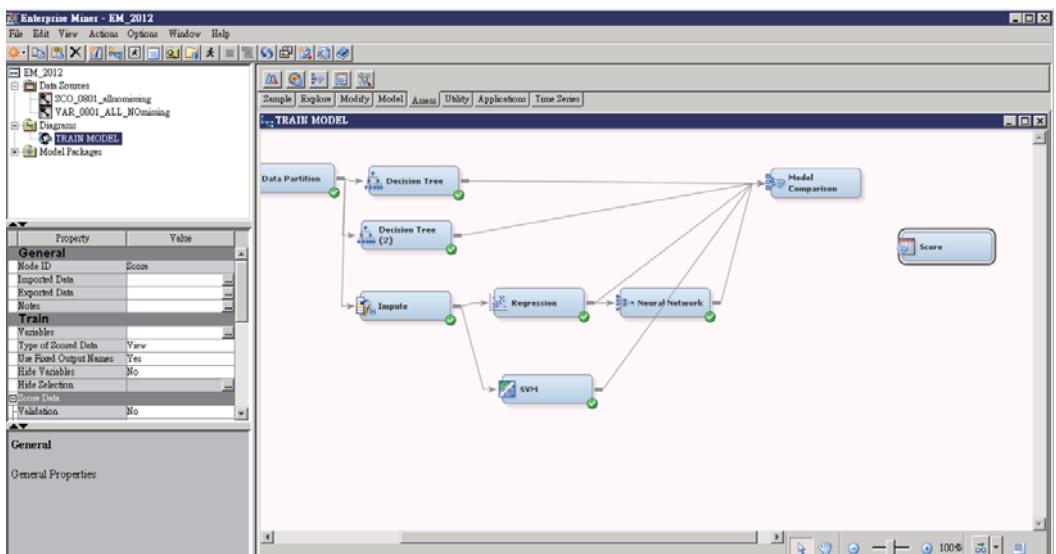


Step3. 此為 Model Comparison(模型比較)節點輸出結果，其中 ROC Chart 為判斷模型好壞程度的重要依據。

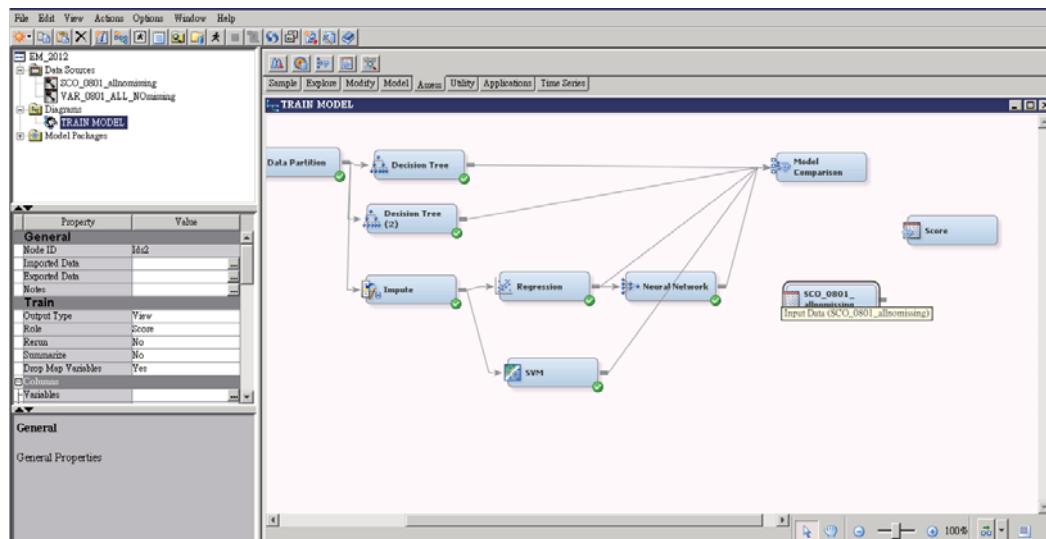
## A2-3 評分

目的：即開始預測未來資料的步驟，建立模型的目的即是以訓練過去的資料而達到預測未來資料的目的。

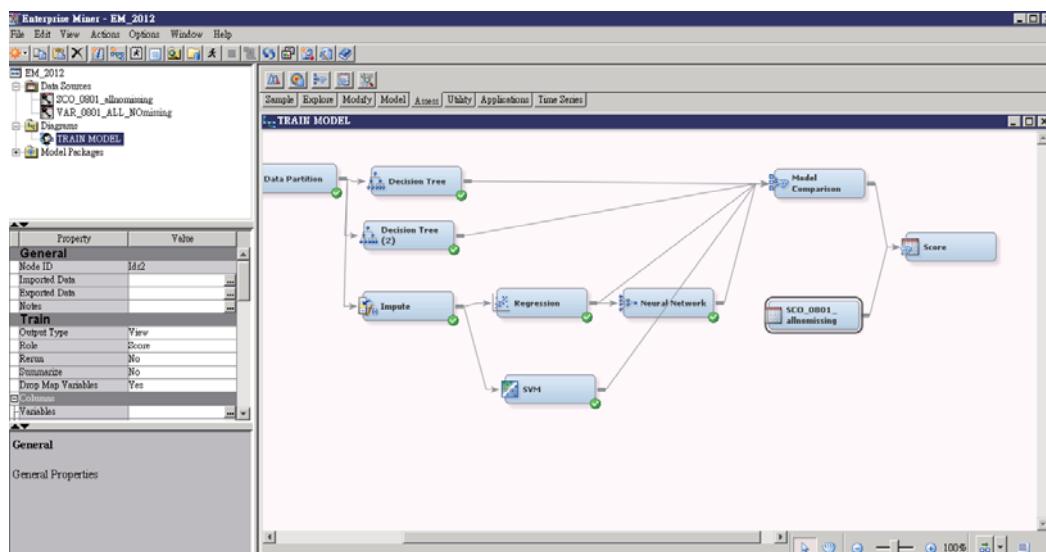
### 3.1 評分步驟



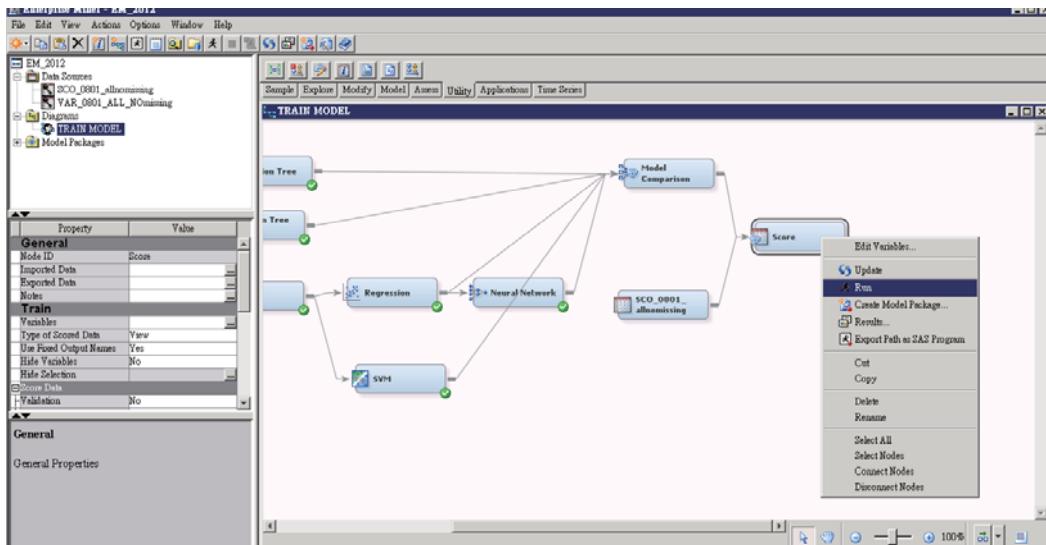
**Step1.** 在 EG 資料表的建置中，已將 SAS 公司與玉山銀行所提供的評分資料整合成資料表以進行銷售顧客之預測，並在先前模型建置前置作業已將評分資料表作匯入資料來源的動作，此時可直接將 Access→Score 節點拖曳至圖表中。



**Step2.** 接著即可將左邊目錄視窗上的 Diagram→SCO\_0801\_allnomissing 資料表直接拉進圖表中。

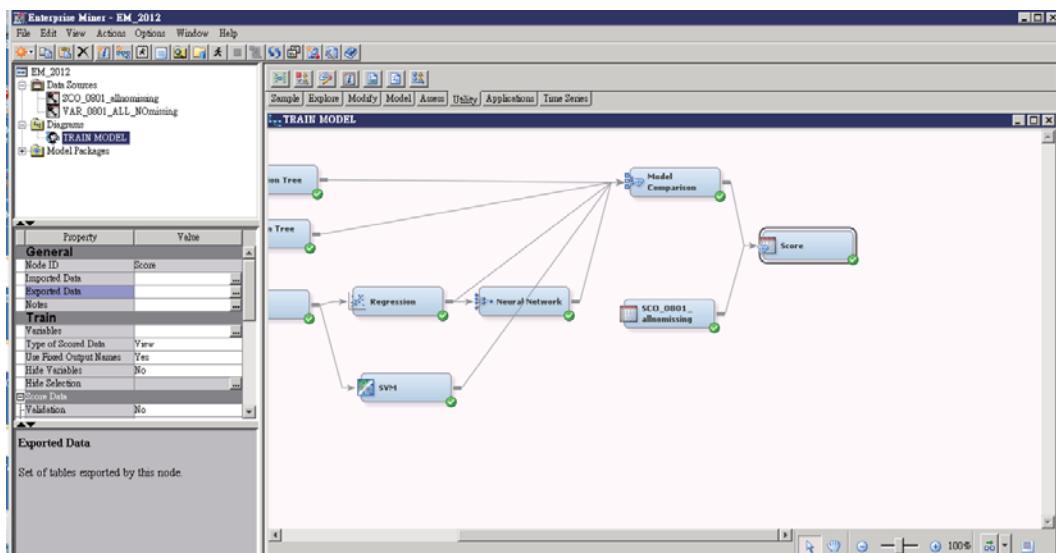


**Step3.** 將 SCO\_0801\_allnomissing 資料表與 Model Comparison(模型比較)節點與 Score 節點做連接。

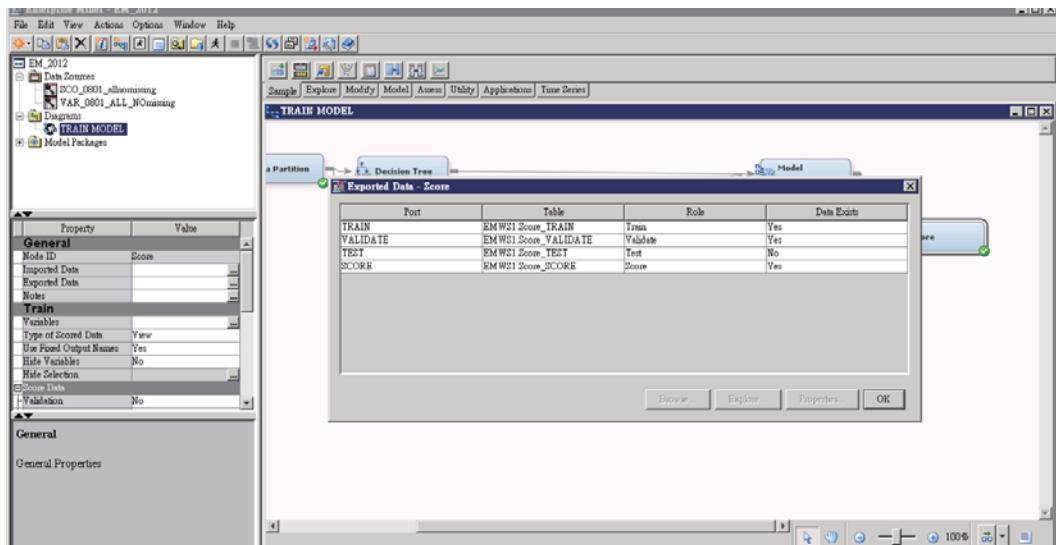


Step4. 在 Score 節點按右鍵→Run。

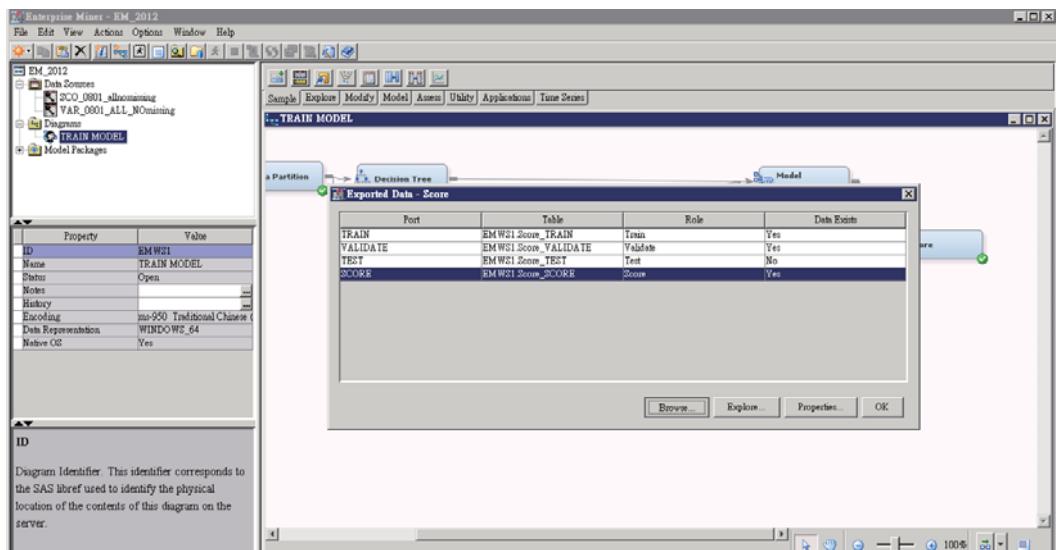
### 3.2 評分結果



Step1. 上步驟執行完畢之後，點選左邊屬性視窗中的 General→Exported Data 可看輸出結果。



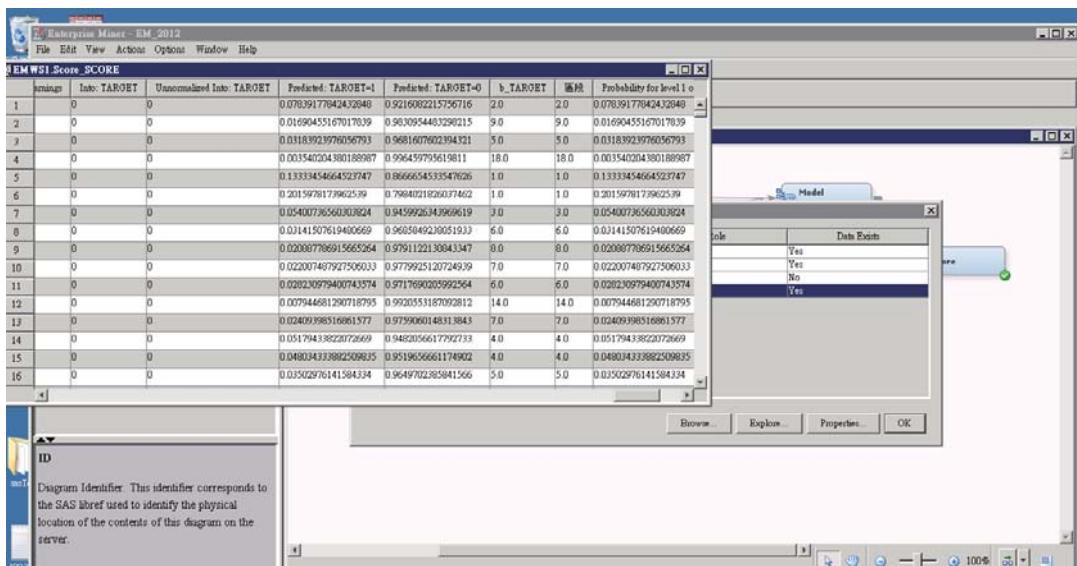
**Step2.** Exported Data 此功能，可以看 TRAIN、VALIDATE、TEST 和 SCORE 之結果。



**Step3.** 同上步驟，本專案欲知道 Score 節點結果，因此點選 Score 進行瀏覽。

ID	性别	子女数	教育程度	婚姻状态	工作年数	年收入	精英	是否拥有财卡	持有正卡数	理财财卡数	本行分期消费标记	本行分期生息标记	本行垫款账户标记	本行逾期账户标记
1	0004931303116997365946720120101M	0.0	8	2	0.0	426568.0	C06	0	2.0	0.0	0	0	0	0
2	00049323039727959323912720120401F	0.0	4	1	0.0	839992.0	C15	0	2.0	0.0	0	0	0	1
3	00044313131153980333746720120101M	0.0	2	2	2.0	480492.0	C04	1	3.0	2.0	1	0	0	1
4	00049313031013971333185120120101M	0.0	4	1	0.0	409996.0	C10	0	1.0	0.0	0	0	0	1
5	00049313031273994355341620120101M	0.0	3	2	1.0	293328.0	C12	0	6.0	0.0	0	0	0	0
6	00049313032093986385009420120101M	0.0	5	2	1.66	180000.0	C05	0	2.0	0.0	0	0	0	0
7	0004931303212963393644020120101M	0.0	3	1	0.0	951996.0	C12	1	9.0	2.0	0	0	0	1
8	00049313032183983365530020120101M	0.0	3	2	0.0	319992.0	C15	0	1.0	0.0	0	0	0	0
9	0004931303223983365704420120101M	0.0	5	1	5.0	399996.0	C12	0	1.0	0.0	0	0	0	0
10	0004931303253985343657520120101M	0.0	3	2	0.33	0.0	C02	0	2.0	0.0	0	0	0	0
11	000493130332398534597220120401M	0.0	4	1	0.0	100000.0	C06	0	1.0	0.0	0	0	0	0
12	00049313033243983353165420120101M	0.0	3	2	1.41	564996.0	C02	0	1.0	0.0	0	0	0	1
13	0004931303404398339400830120401M	0.0	3	2	0.0	322692.0	C15	0	1.0	0.0	0	0	0	0
14	0004931303404398339400830120101M	0.0	1	2	2.5	400492.0	C17	0	1.0	0.0	0	0	0	0
15	0004931303404398339400830120401M	0.0	1	2	2.5	480492.0	C17	0	1.0	0.0	0	0	0	0
16	0004931303417398336301960120101M	0.0	3	2	0.66	331200.0	C06	0	1.0	0.0	0	0	0	1
17	00049313034213986363695520120101M	0.0	5	1	4.0	159900.0	C10	0	2.0	0.0	0	0	0	0
18	0004931303506398437739420120101M	0.0	4	2	0.0	259992.0	C14	0	1.0	0.0	1	0	1	1
19	00049313035143984383564220120101M	0.0	5	2	1.0	259992.0	C02	0	2.0	0.0	0	0	0	0
20	0004931303516399637147920120101M	0.0	3	2	0.0	199992.0	C17	0	2.0	0.0	1	0	1	1
21	000493130360739833938104020120401M	0.0	5	2	9.0	1165980.0	C10	0	1.0	0.0	0	0	0	1
22	00049313036113985355171620120101M	0.0	5	2	6.0	180000.0	C15	0	1.0	0.0	0	0	0	0
23	0004931303703982313262020120101M	0.0	2	2	0.16	159990.0	C05	0	2.0	0.0	0	0	0	1
24	00049313037073983323206020120101M	0.0	3	2	1.25	199992.0	C12	0	2.0	0.0	0	1	0	0
25	0004931303707398336382330120101M	0.0	4	2	0.0	986798.0	C02	0	1.0	0.0	0	0	0	1
26	00049313037253984313521350120101M	0.0	5	1	2.5	120000.0	C02	0	1.0	0.0	1	0	1	0
27	00049313038043984392244720120101M	0.0	3	2	4.0	399996.0	C10	0	1.0	0.0	0	0	0	0
28	00049313038103982375468220120101M	0.0	8	1	0.0	322692.0	C15	0	1.0	0.0	0	0	1	1
29	0004931303813399053658031720120101M	0.0	3	2	0.0	345996.0	C15	0	3.0	0.0	0	0	1	0

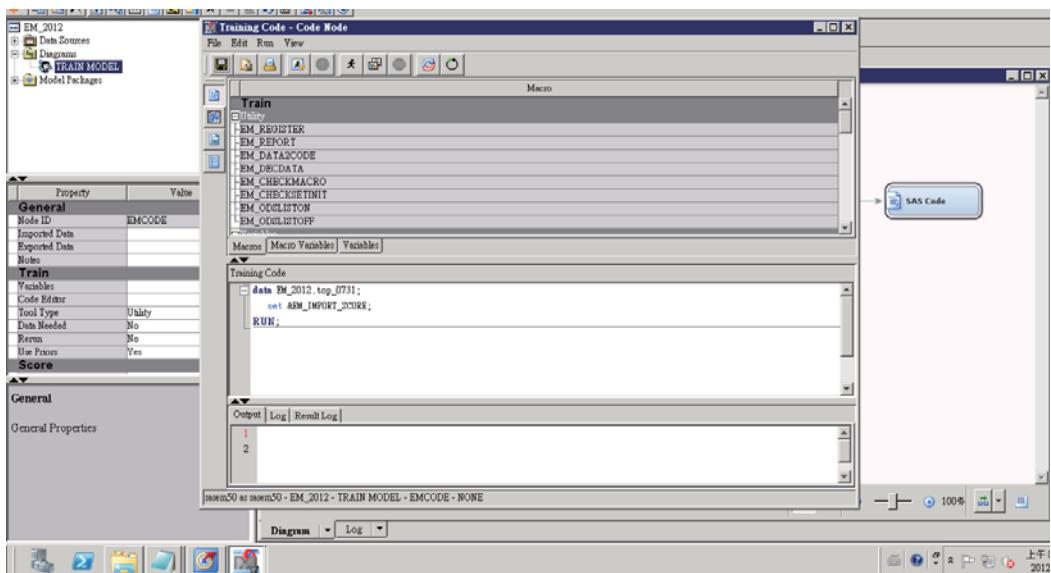
Step4. 此為瀏覽後全部的資料之部分截圖。



Step5. 可觀察 TARGET 為 1 之機率。

**Step6.** 同上步驟，亦可觀察 Target=1 的個數。

**Step8.** 最後須將 SAS 資料匯出，因此須在圖表中拖曳進來一個 SAS Code 節點，藉由 SAS Code 將資料匯出。節點位置於 Utility→SAS Code 將節點拖曳至圖表中，接著選擇 Train→Code Editor 即出現此視窗。



### Step9. 接著輸入程式碼

Data EM\_2012.top\_0731

EM\_2012 為 Library 名稱，top\_0731 為匯出之檔案名稱，輸入完即可進行儲存。儲存完後即可在 SAS Code 節點按右鍵→Run。

The screenshot shows the SAS Results window with the 'SAS Code' mode selected. The 'Output' tab displays the log file content:

```

1 -----
2 使用者:          sasem50
3 日期:            2012年08月02日
4 時間:            03時58分42秒
5 -----
6 * 1個輸出
7 -----
8
9
10
11
12 變數摘要
13
14 角色      量值屬性 次數計數
15
16 ASSESS    BINARY     1
17 ASSESS    NOMINAL   1
18 CLASSIFICATION NOMINAL  3
19 ID        INTERVAL   1
20 ID        NOMINAL   1
21 INPUT     BINARY     1
22 INPUT     INTERVAL   13
23 INPUT     NOMINAL   4
24 PREDICT   INTERVAL   2
25 REJECTED  BINARY     13
26 REJECTED  INTERVAL   55
27 REJECTED  NOMINAL   2
28 RESIDUAL  INTERVAL   2
29 SEGMENT   NOMINAL   2
30 TARGET    BINARY     1
31
32

```

Step10. 此為執行完後之 log 檔案