SAS校園資料採礦競賽

IMTKU 淡江資管隊

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)

大綱(Outline)

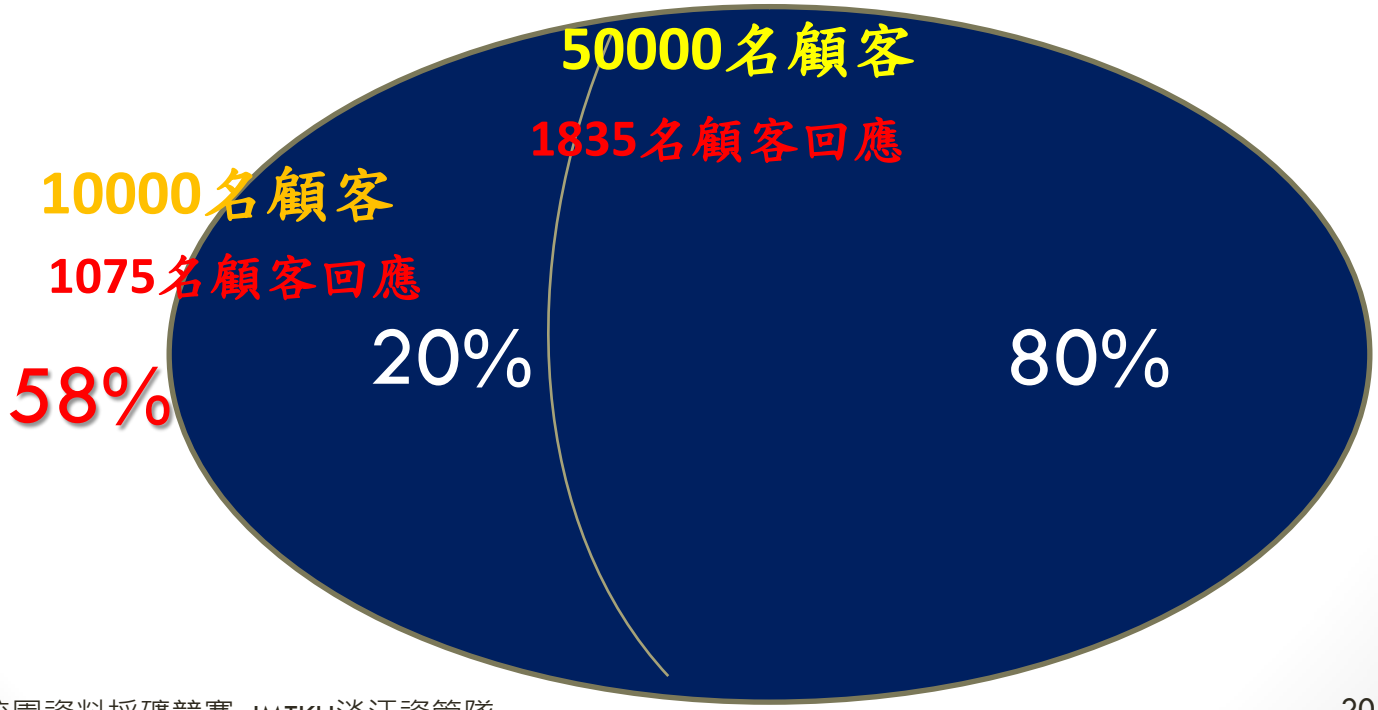
- 管理摘要(Executive Summary)
- 模型建置(Model Development)
- 研究結果(Experimental Results and Discussions)
- 結論(Conclusion)

為何作這項研究?
(Why this research?)

- 有效運用模型技術，找出回應率高的潛在顧客名單
- 以同樣的行銷成本，創造玉山產品電話行銷之最大利潤

研究中發現什麼?(What was found)

- 由玉山銀行所提供的10萬筆顧客名單作為訓練資料集，並結合SAS Enterprise Miner(EM)搭配SAS Enterprise Guide(EG)建置模型預測5萬名顧客名單，本團隊提供一萬名建議行銷顧客名單，並與主辦單位提供之五萬名正確回應顧客名單做比較，本團隊之模組預測回應率高達10.75%



研究發現所代表的意義?
(What those findings mean)

假設：

- 平均行銷一名顧客的電話成本為10元。
- 客戶回應即成為玉山顧客(回應後不一定成為玉山顧客，有可能因部分產品條件而不購買玉山產品或是玉山可能會因為信用條件而婉拒該顧客，本專案省略此部分)
- 本專案僅考慮行銷時的電話成本，其餘直接成本與間接成本暫不考慮。

研究發現所代表的意義?
(What those findings mean)

50000名顧客

回應顧客1835人

每通電話行銷成本10元

平均成功一位顧客成本所需

10/3.67%=272元

總行銷成本為50萬元

10000名顧客

回應顧客1075人

每通電話行銷成本10元

平均成功一位顧客成本所需

10/10.75%=93元

總行銷成本為10萬元

成本節省：(272-93)/272*100%=65%

研究發現所代表的意義?
(What those findings mean)

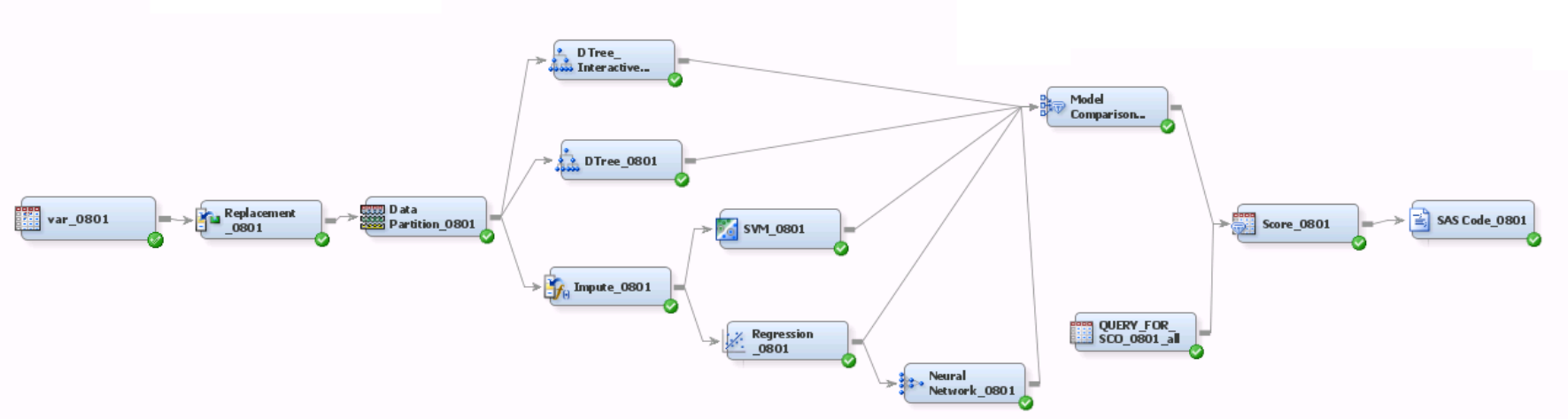
- 透過電話產品行銷，結合我們所建置的模型之回應顧客建議銷售名單，可以替公司省下65%之行銷成本

所採取之行動?(What action)

- 針對模型所挑出的顧客行銷名單，配合電話行銷人力規劃及產品設計，分批進行行銷專案。
- 可設計行銷頻率，例如：每半年重複一次，在不影響顧客觀感下，可提高名單的利用價值。

模型建置(Model Development)

模型建置-TIME WINDOW

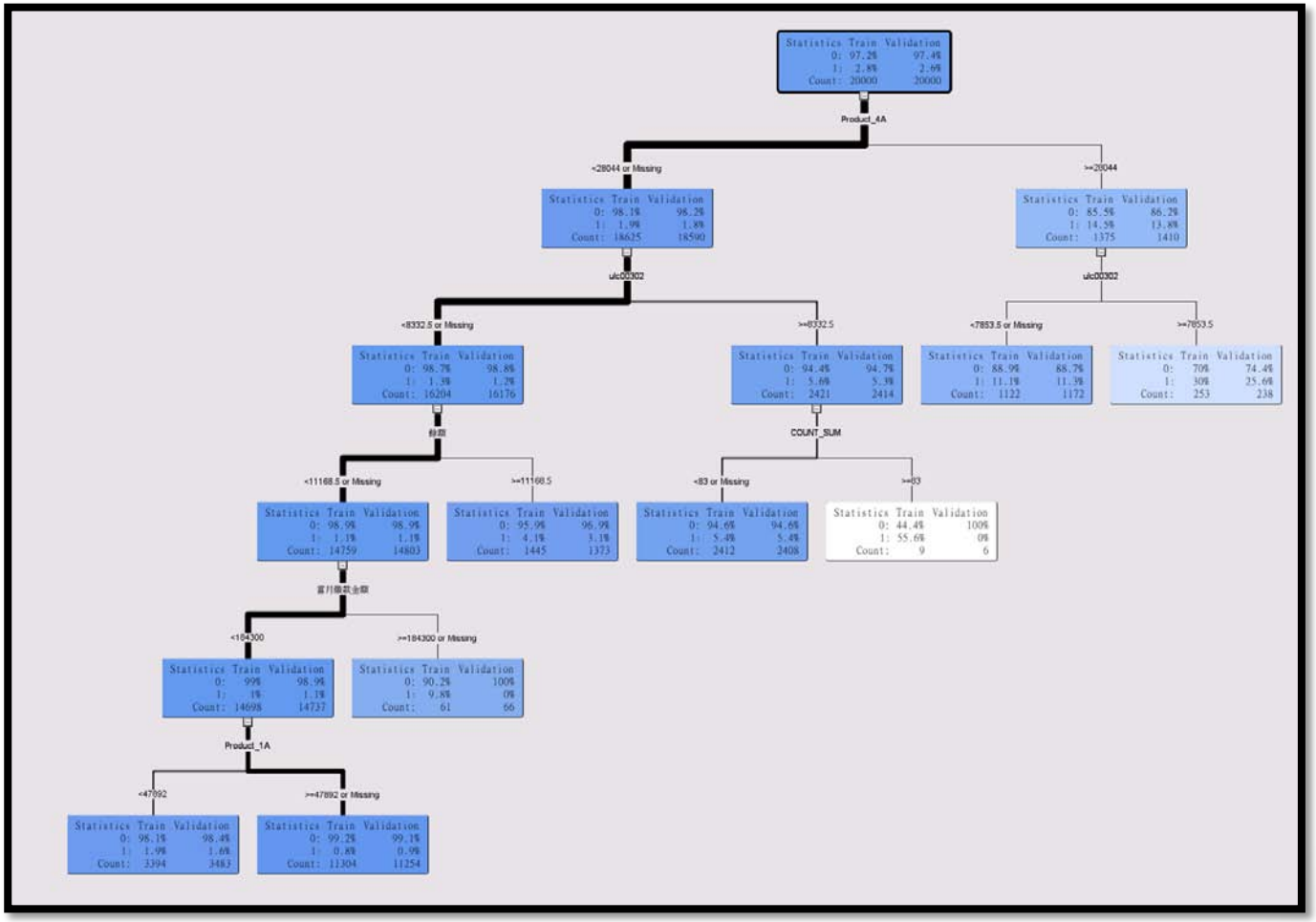


圖一、模型建置之Time Window

模型建置(Model Development)

- 決策樹(Decision Tree)
- 迴歸(Regression)
- 類神經(Neural Network)
- 支持向量機(SVM)
- 模型比較(Model Comparison)

模型建立(EM)-決策樹(Decision Tree)



圖二、決策樹模型

模型建立(EM)-決策樹(Decision Tree)

- 所分枝的變數依序是

  1.PRODUCT_4A(信貸正常)

  2.ULC00302(全體金融機構預借現金金額)

  3. ULC00303(全體金融機構上期未繳金額(循環)-截至2011年7月之未繳金額餘額)

  4. CST_018(當月繳款金額)

- 其中前三個變數屬於衍生變數，作為決策樹分枝的依據可提升模型之預測能力

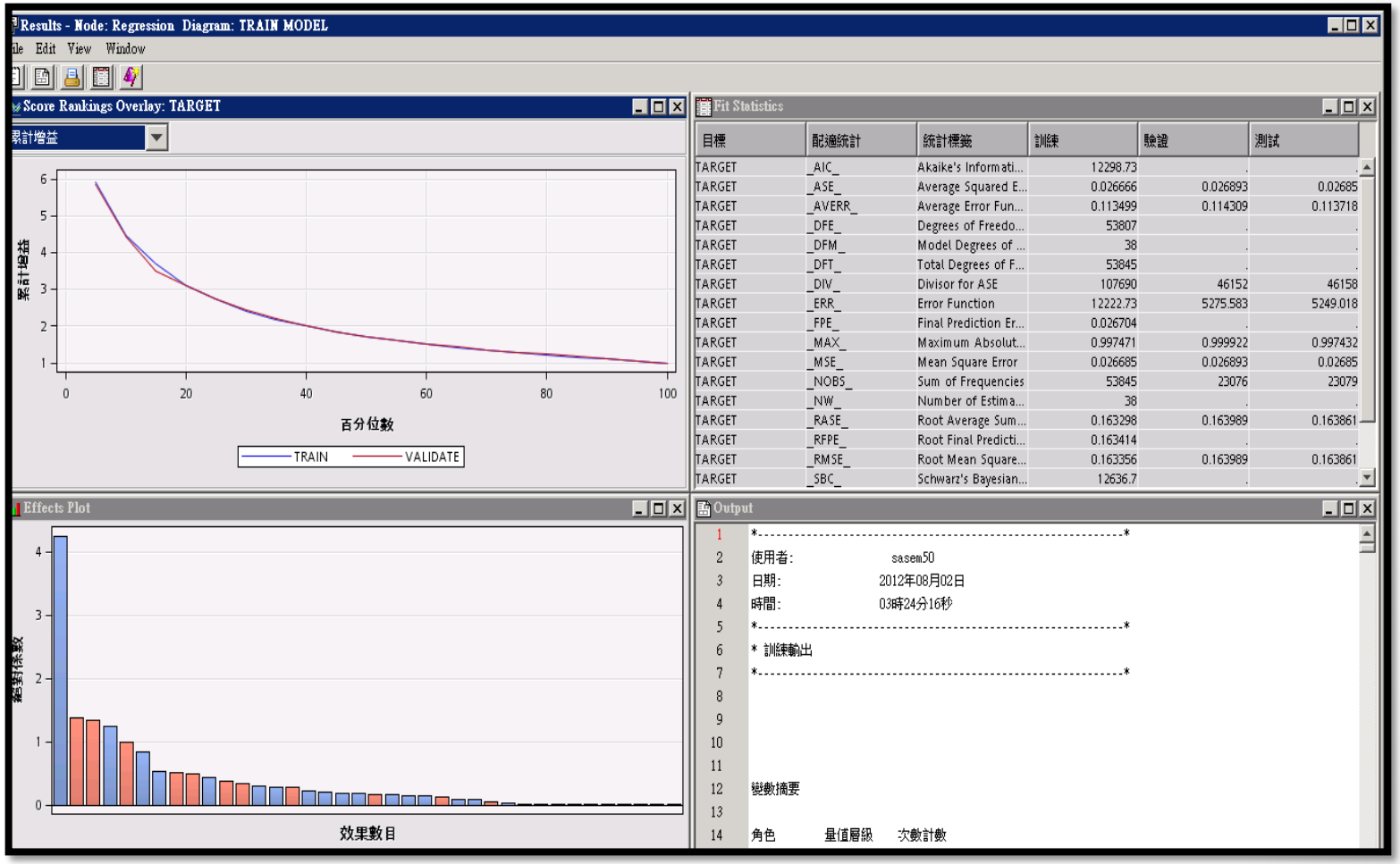
模型建立(EM)-迴歸(Regression)

Property	Value
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No

Property	Value
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No

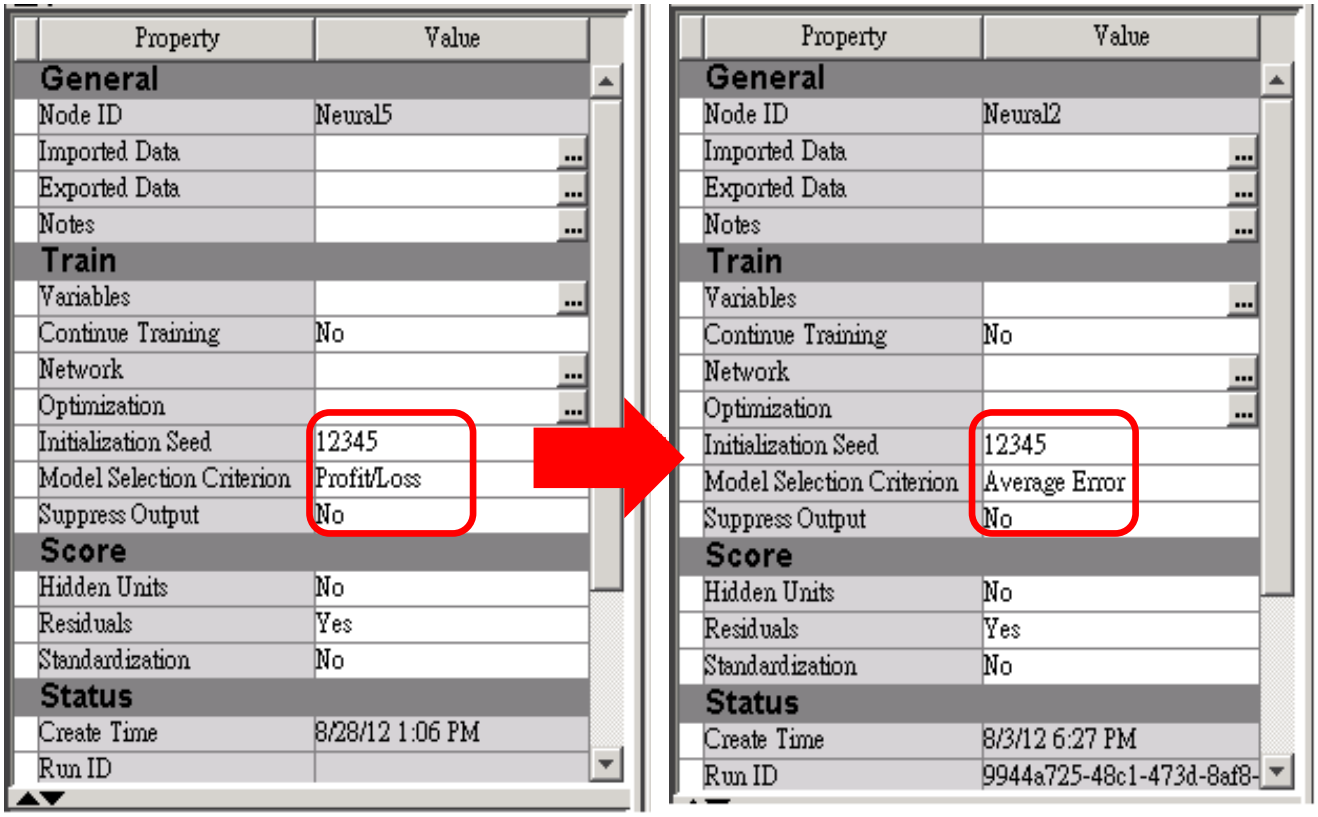
圖三、迴歸參數調整

模型建立(EM)-迴歸(Regression)



圖四、迴歸模式輸出結果

模型建立(EM)-類神經網路(Neural Network)

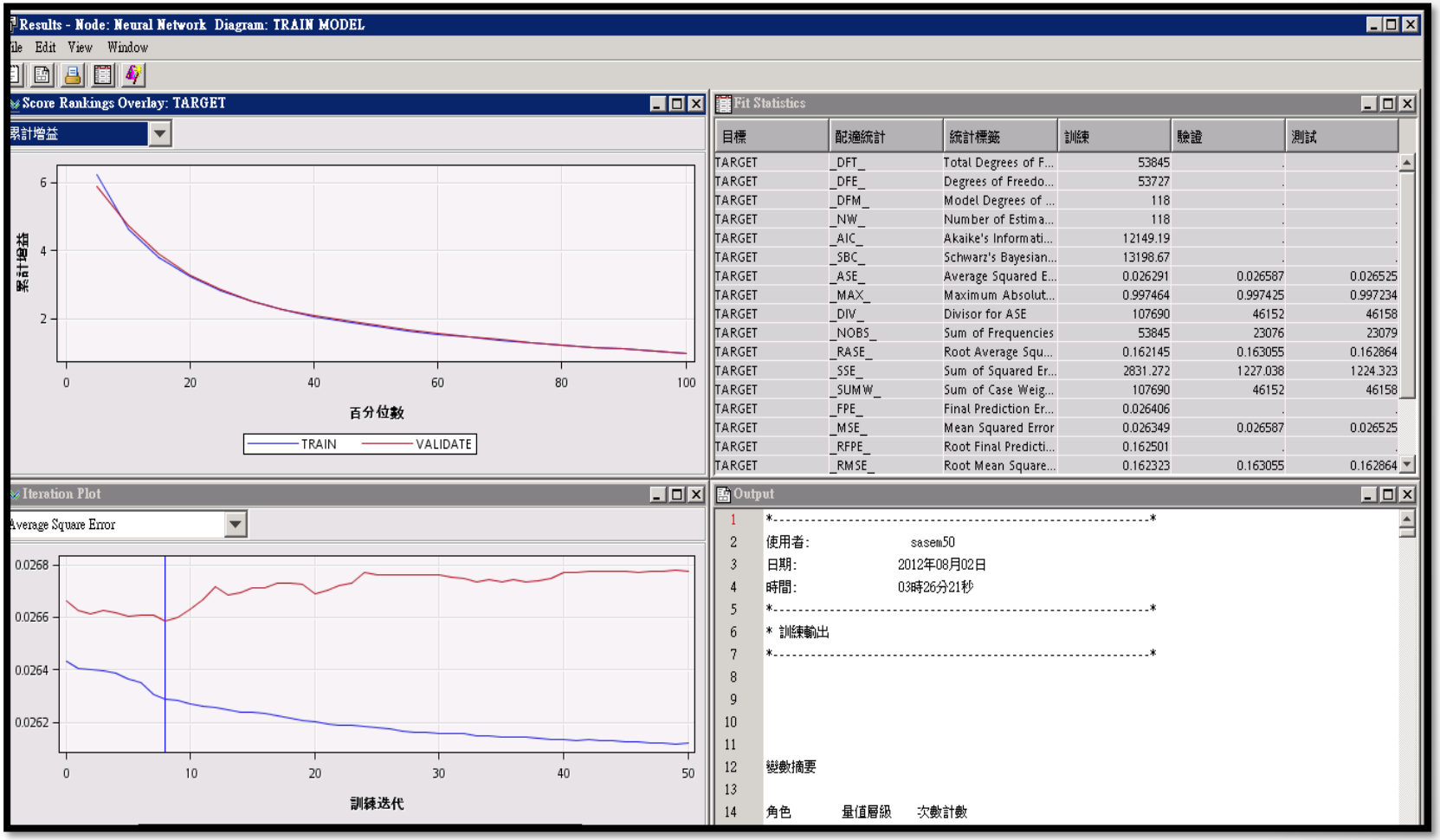


Property	Value
General	
Node ID	Neural5
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	8/28/12 1:06 PM
Run ID	

Property	Value
General	
Node ID	Neural2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	8/3/12 6:27 PM
Run ID	9944a725-48c1-473d-8af8-

圖五、類神經網路參數調整

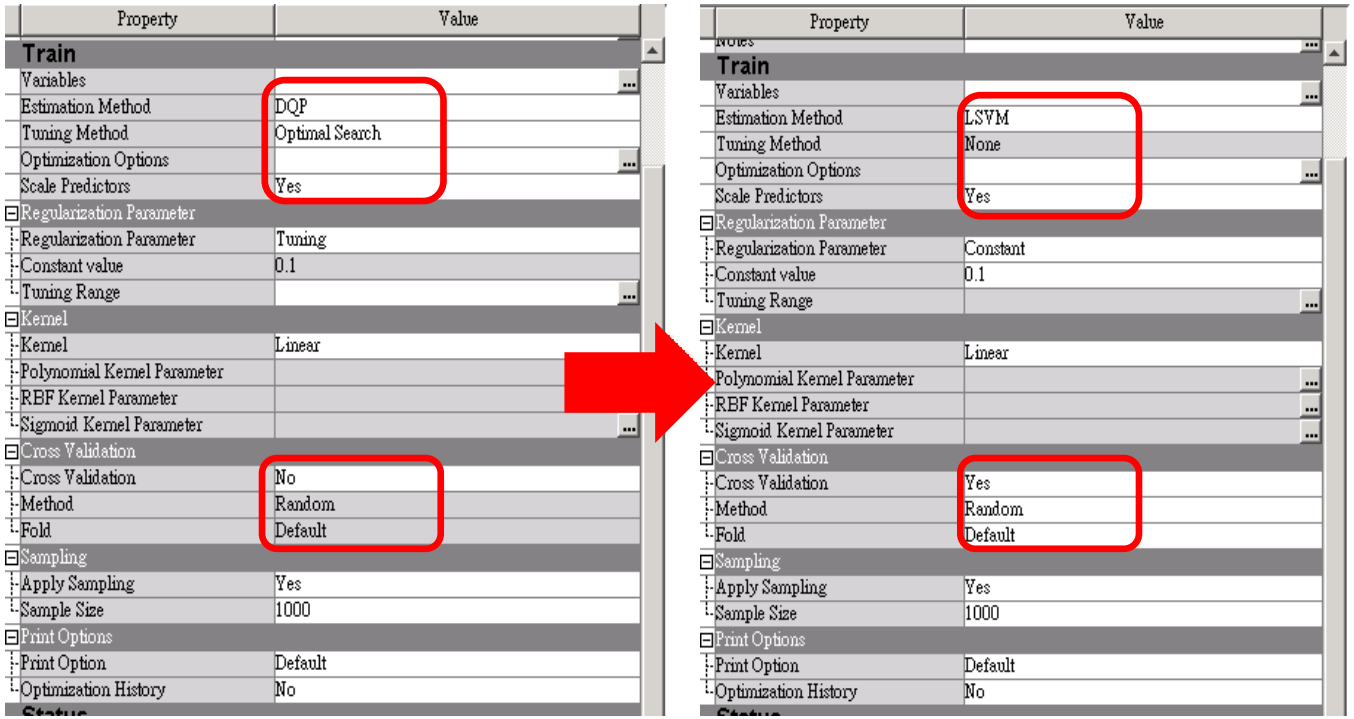
模型建立(EM)-類神經網路(Neural Network)



目標	配適統計	統計標籤	訓練	驗證	測試
TARGET	_DFT_	Total Degrees of F...	53845	.	.
TARGET	_DFE_	Degrees of Freedo...	53727	.	.
TARGET	_DFM_	Model Degrees of ...	118	.	.
TARGET	_NW_	Number of Estima...	118	.	.
TARGET	_AIC_	Akaike's Informati...	12149.19	.	.
TARGET	_SBC_	Schwarz's Bayesian...	13198.67	.	.
TARGET	_ASE_	Average Squared E...	0.026291	0.026587	0.026525
TARGET	_MAX_	Maximum Absolut...	0.997464	0.997425	0.997234
TARGET	_DIV_	Divisor for ASE	107690	46152	46158
TARGET	_NOBS_	Sum of Frequencies	53845	23076	23079
TARGET	_RASE_	Root Average Squ...	0.162145	0.163055	0.162864
TARGET	_SSE_	Sum of Squared Er...	2831.272	1227.038	1224.323
TARGET	_SUMW_	Sum of Case Weig...	107690	46152	46158
TARGET	_FPE_	Final Prediction Er...	0.026406	.	.
TARGET	_MSE_	Mean Squared Error	0.026349	0.026587	0.026525
TARGET	_RFPE_	Root Final Predicti...	0.162501	.	.
TARGET	_RMSE_	Root Mean Square...	0.162323	0.163055	0.162864

圖六、類神經網路輸出結果

模型建立(EM)-支持向量機(SVM)

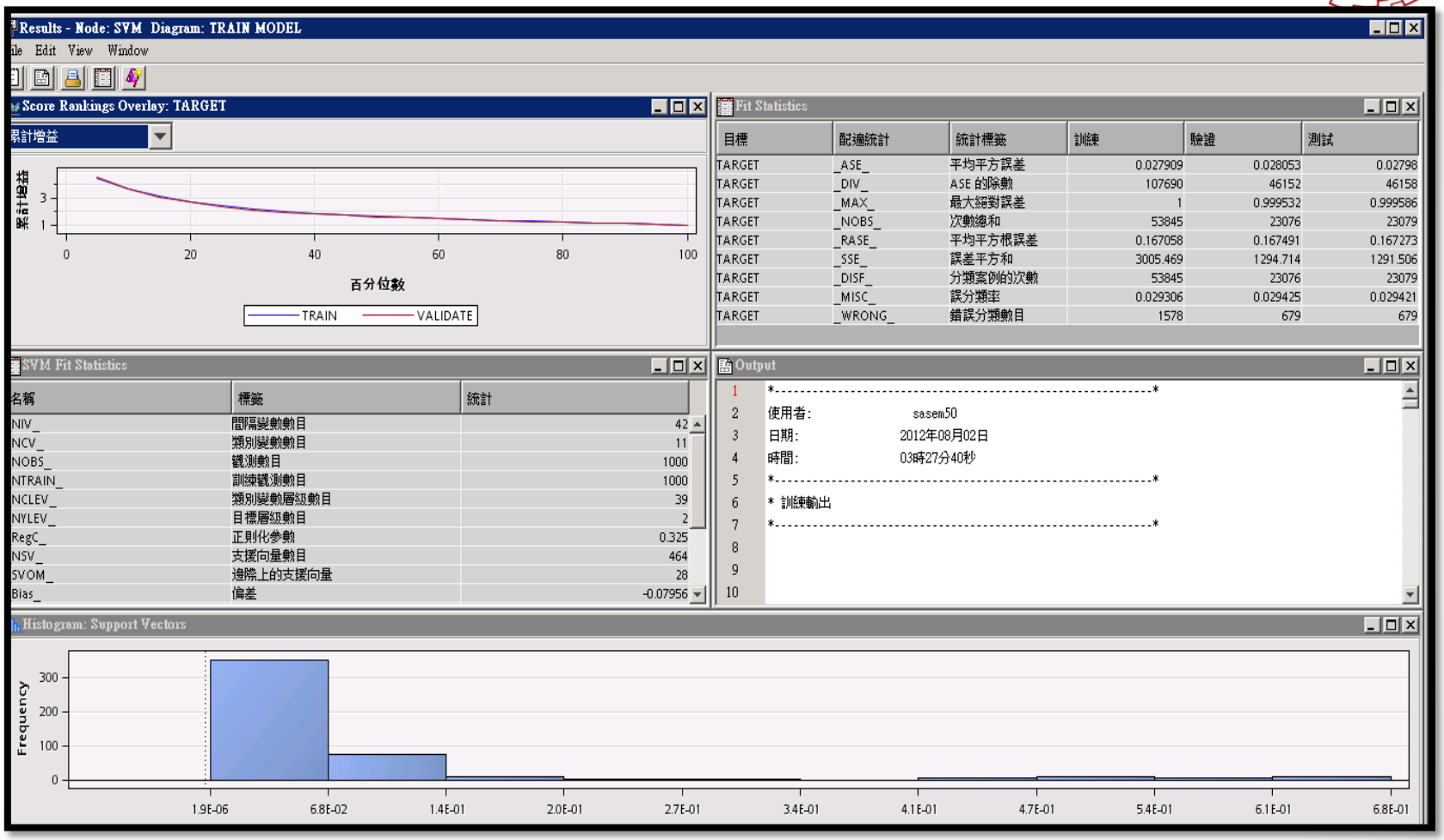


Property	Value
Train	
Variables	
Estimation Method	DQP
Tuning Method	Optimal Search
Optimization Options	
Scale Predictors	Yes
Regularization Parameter	
Regularization Parameter	Tuning
Constant value	0.1
Tuning Range	
Kernel	
Kernel	Linear
Polynomial Kernel Parameter	
RBF Kernel Parameter	
Sigmoid Kernel 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
Train	
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

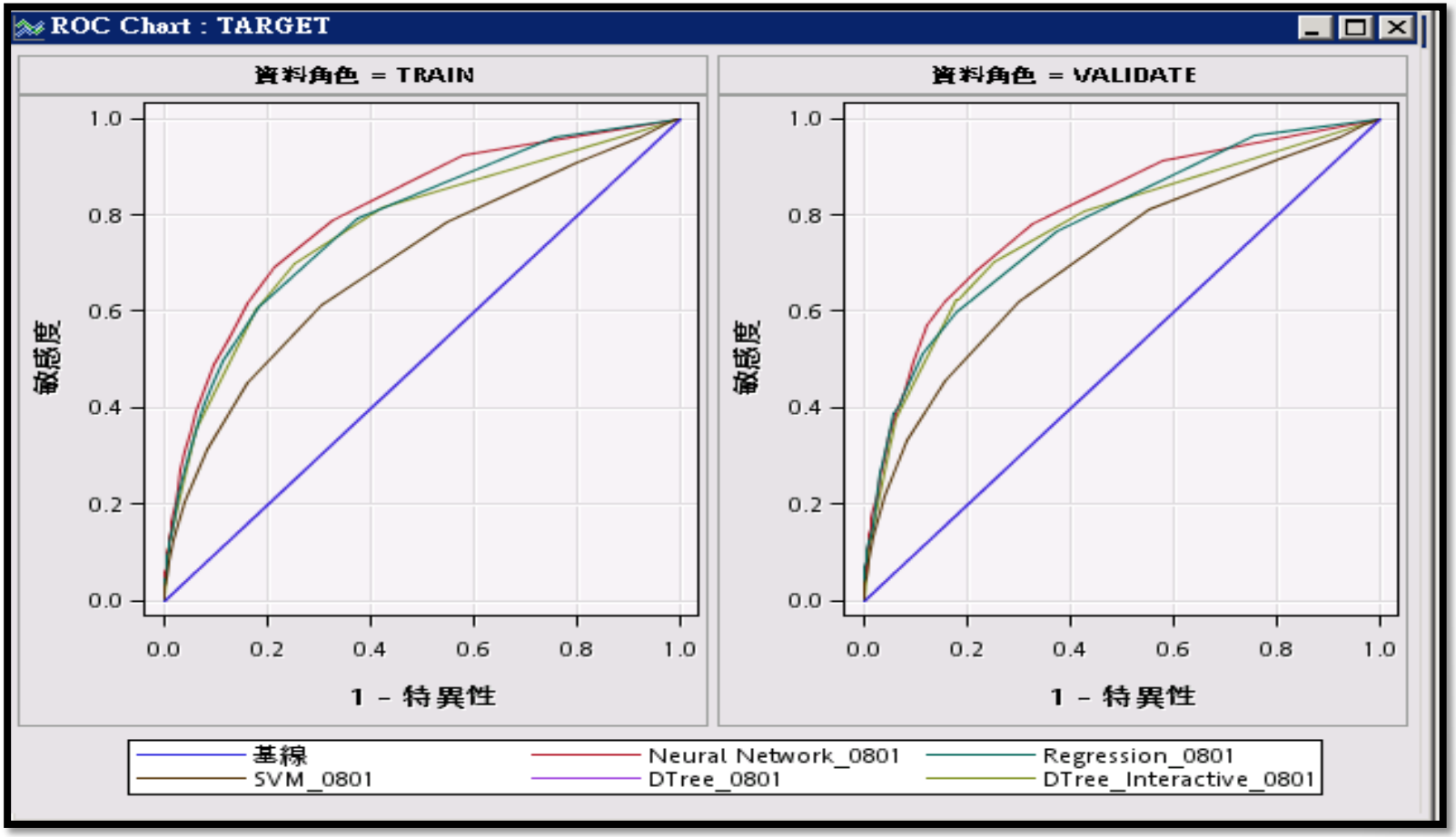
圖七、支持向量機(SVM)參數調整

模型建立(EM)-支持向量機(SVM)



圖八、支持向量機(SVM)輸出結果

模型建立(EM)-模型比較(Model Comparison)



圖九、模型比較-輸出結果(ROC)

模型建立(EM)-模型比較(Model Comparison)

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

圖十、模型比較-輸出結果(LIFT)

研究結果(Experimental Results)

- **資料集與模型比較與結果**

本研究依據不同變數反覆測試，整理出3筆資料集SAS_0731 (54個變數)、SAS_0801 (78個變數)以及SAS_0803 (108個變數)，皆運用相同之模組與方法做行銷建議預測。各資料集之差異比較如下表：

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**半年**資料	將月份整合為**半年**資料	將每個月份**分開呈現**(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**半年**資料	將月份整合為**半年**資料	將每個月份**分開呈現**(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

圖十一、三筆資料集差異比較表

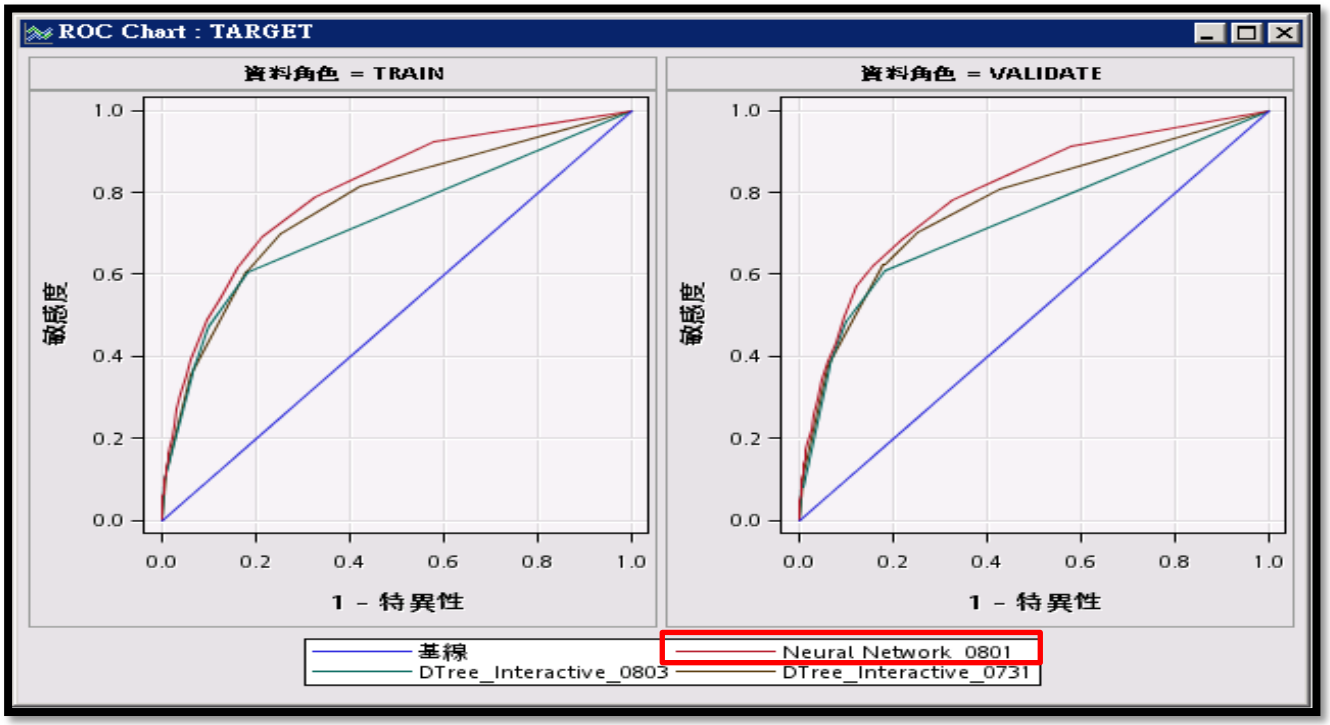
研究結果(Experimental Results)

各資料表執行預測後之結果，其TRAIN累積增益數據比較如下表：

MODEL\DATA SET	SAS_0731 (54個變數)	SAS_0801 (78個變數)	SAS_0803 (108個變數)
Regression (迴歸)	3.36	4.38	3.33
Decision Tree-1 (決策樹-手動自建樹)	4.09	4.09	4.41
Neural Network (類神經網路)	3.9	4.63	4.01
Decision Tree (決策樹-自動樹)	1	1	1
SVM	2.49	3.31	2.33

圖十二、三種最佳模型比較表

研究結果(Experimental Results)



圖十三、三種最佳模型比較-輸出結果(ROC)

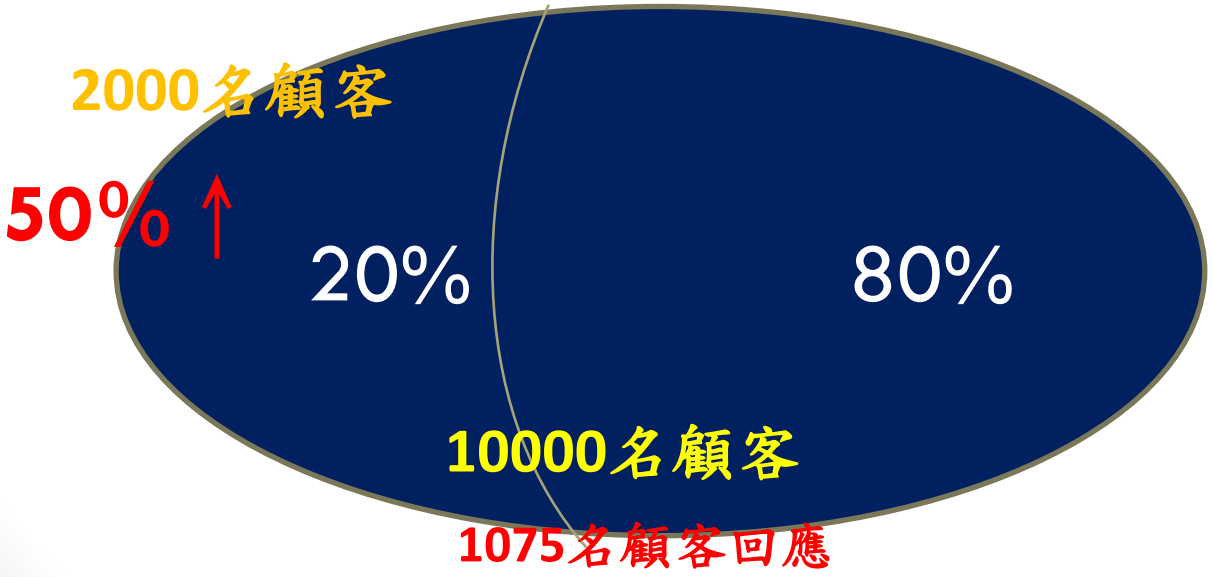
Neural Network_0801即SAS_0801資料集之Neural Network類神經模型ROC Curve為最理想之結果。由以上數據與圖表所示，因此本團隊最後決定使用SAS_0801資料集的Neural Network類神經模型為最後的決策模型。

結論(Conclusion)

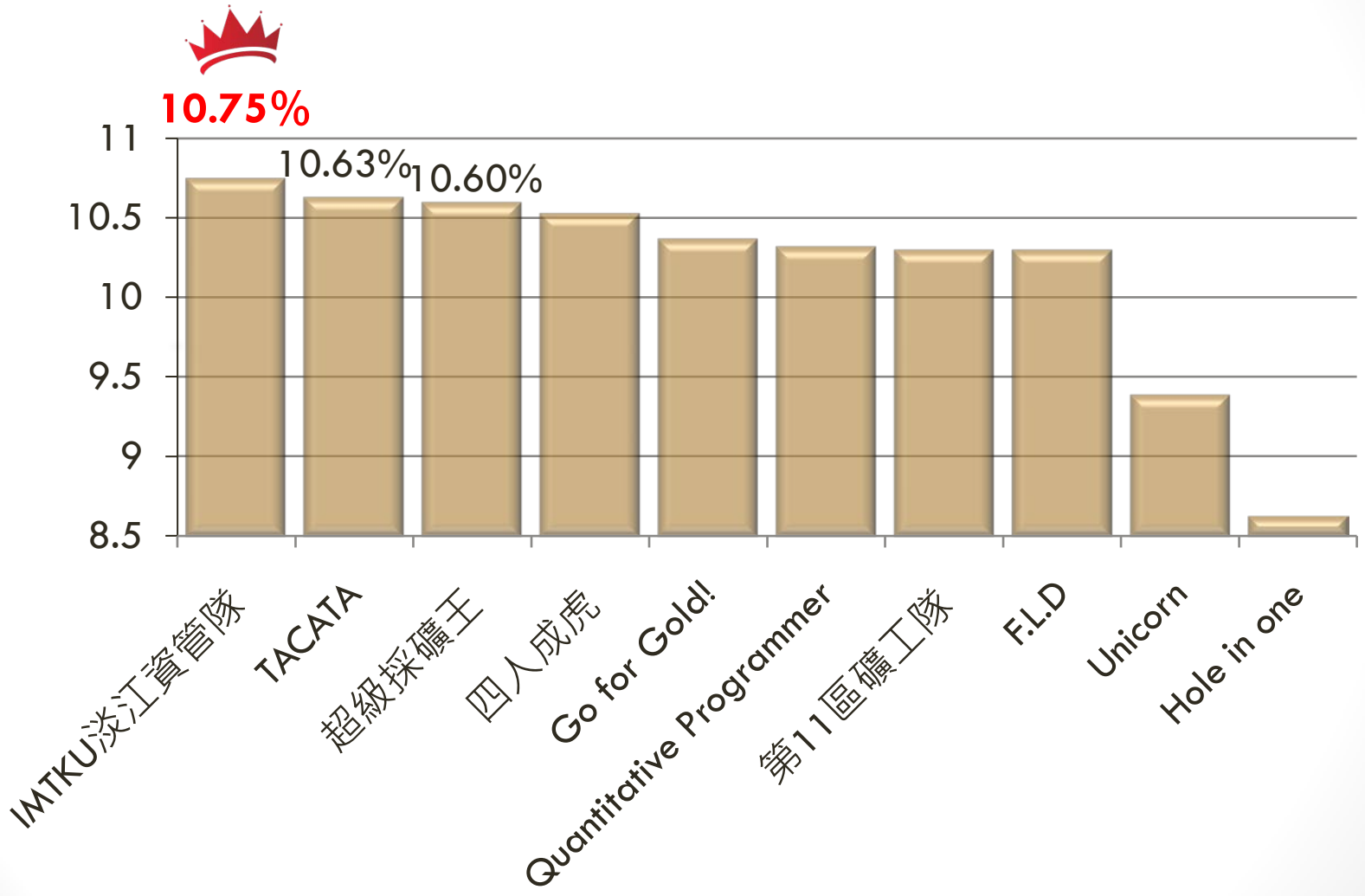
- 經衍生變數評估與比較後，最終運用本小組整合衍伸後之SAS_0801資料集進行預測所提出10000筆建議銷售名單，其成交的預測機率範圍介於約0.70至0.05之間
- 使用SAS_0801資料集的類神經模型於預測出潛在顧客名單之回應率表現比較好
- 經反覆測試發現衍生變數愈多並不表示結果會愈符合預期

未來展望(Future Works)

- 未來運用本團隊之模型進行預測，若能取一萬名顧客以下之顧客即能達到理想結果，意即降低行銷成本，即能達到最佳效益，最終即可不必取一萬名而是取能達到最佳效益的顧客數量。
- 例如:若透過80/20法則，如果只需採用前20%人數作為建議行銷顧客名單(2000名回應顧客)，希望可達到近50%以上之顧客回應率，即能大幅降低行銷成本



前10名隊伍回應率之比較



致謝(Acknowledgment)

- 感謝SAS臺灣提供一個公正而又富挑戰性的舞台的機會，各校資料採礦高手一同組隊挑戰企業實例個案
- 感謝玉山銀行提供機會，將在經營實際所面對的問題以及相關數據資料，提供隊伍進行資料採礦分析，深入了解如何利用所學採礦理論與企業的實際案例做結合
- 感謝SAS的專業顧問及玉山銀行實際操作採礦分析的專家提供請益機會，達到理論與實務的實際結合學習作用。

第一屆SAS校園資料採礦競賽
IMTKU淡江資管隊成員

Q&A

SAS校園資料採礦競賽

IMTKU 淡江資管隊

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)