2019國泰大數據競賽

透過機器學習最小化公司成本

唏哩呼嚕吃不胖

目錄

CONTENTS

PART 1

研究動機

研究動機

目的

- 挖掘出保險需求較高的客戶
- 區分目標顧客群
- 考慮成本下，提供精準的銷售名單

問題

- 資料不齊全
- 購買族群相對稀少

PART 2

資料處理與探索

資料處理與探索 - 補值介紹

Generalized Low Rank Model (GLRM)

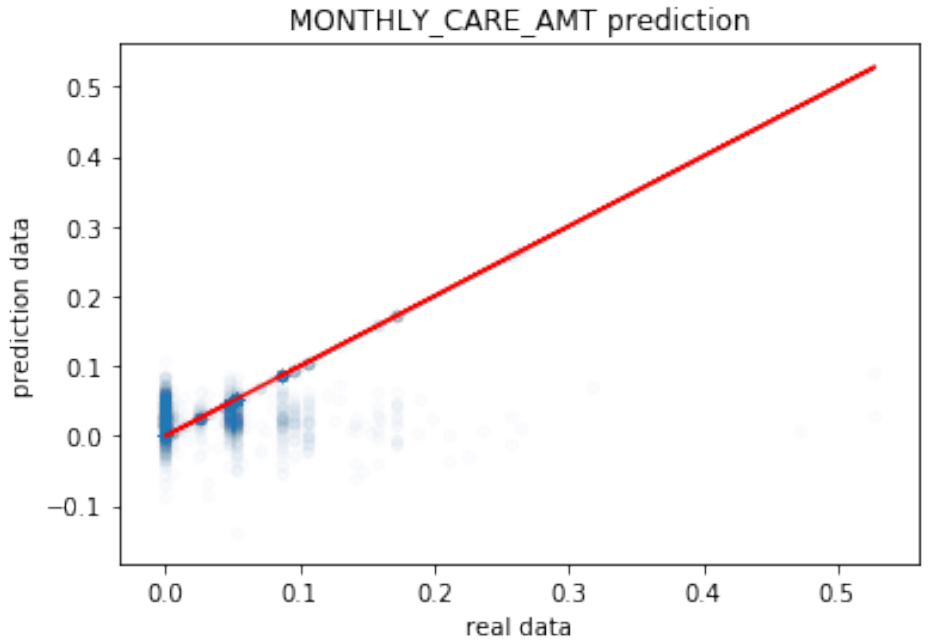
- 未記錄值的推測方式，可應用於推薦系統
- 利用validation data製造遺失值，並進行驗證

Forward Predict

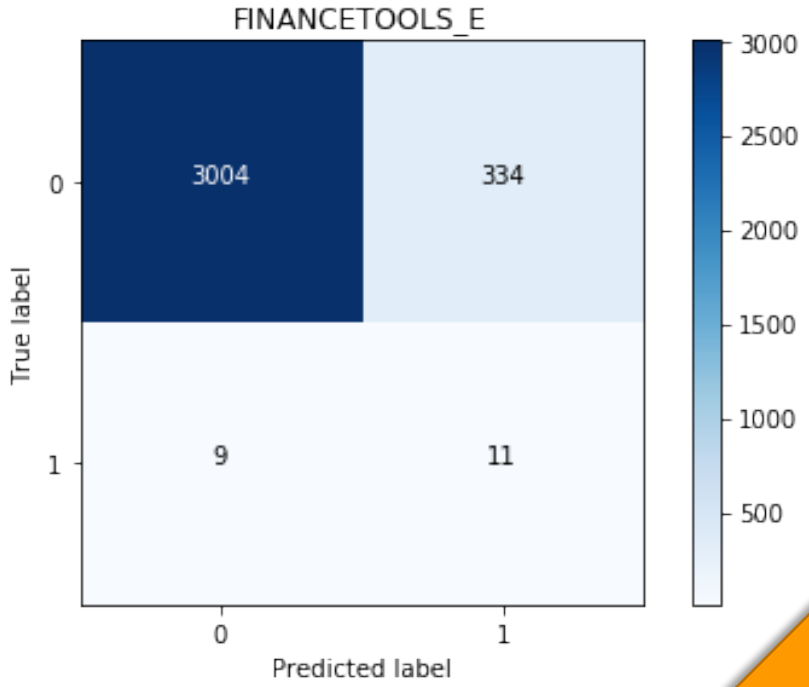
- 針對GLRM補值結果較差類別變數，ACC<0.8
- 抽樣1000組類別變數的排列組合，選取最優者當作補值方式

資料處理與探索 - 補值結果

連續型變數



類別型變數

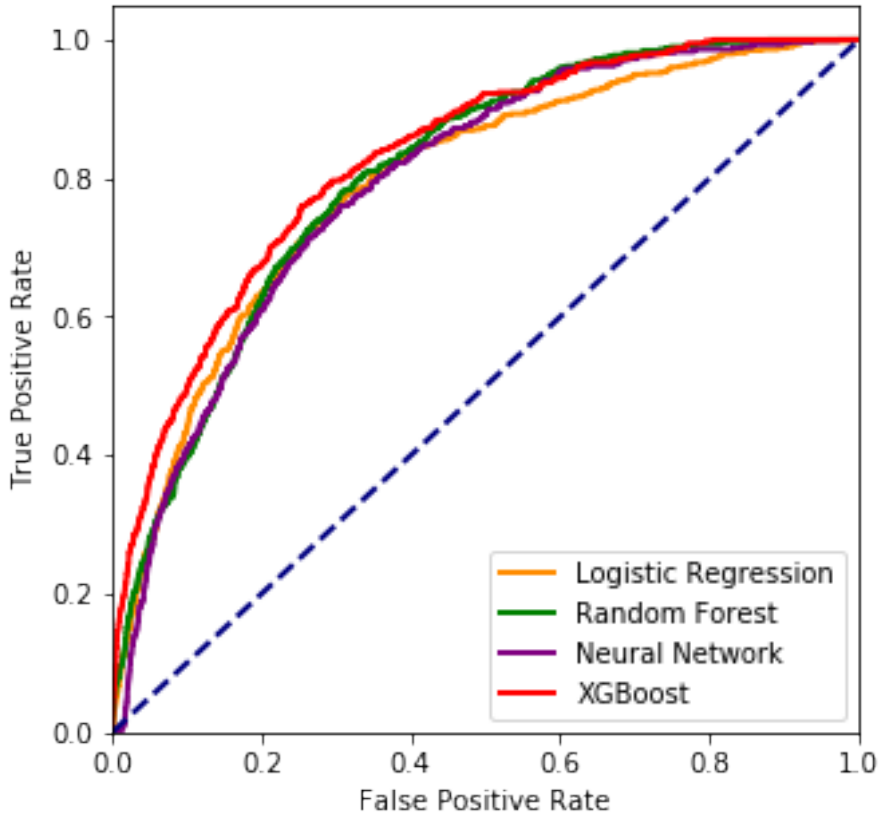


資料處理與探索 – 模型選擇

XGBoost 補值比較

補值方法	預測結果
不補值	0.84407
GLRM	0.84338
GLRM + forward predict （only test）	0.85195

模型ROC curve 比較圖

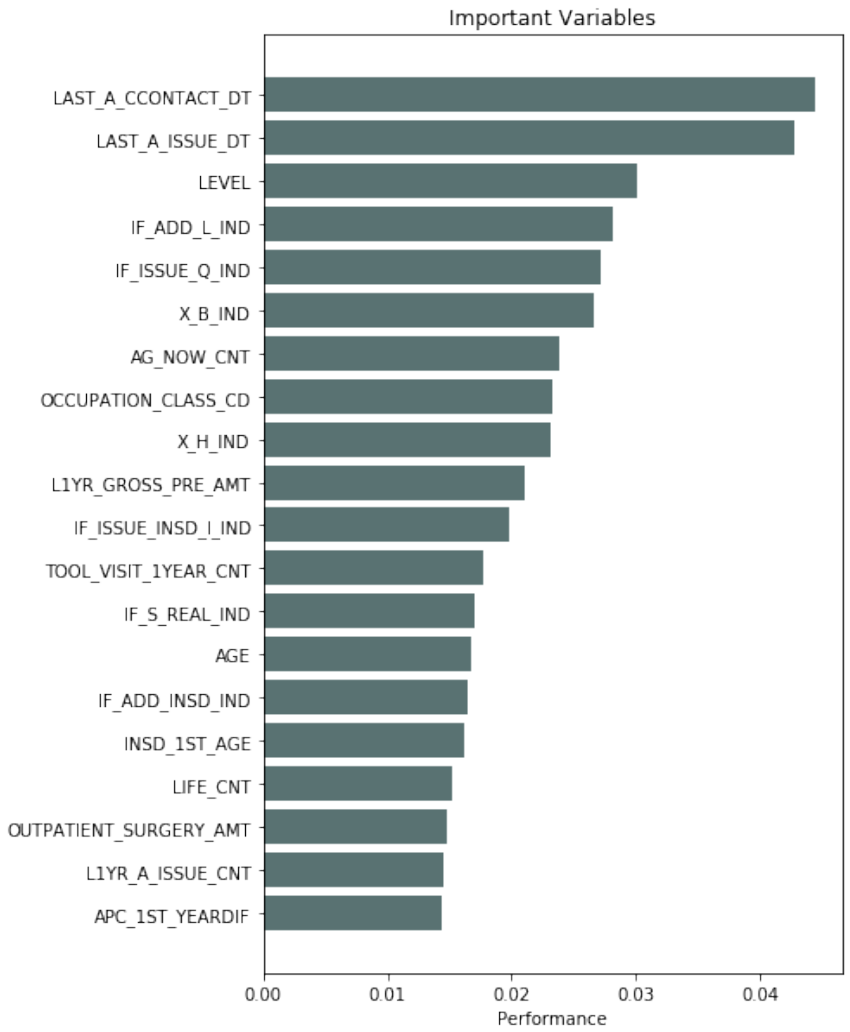


PART 3

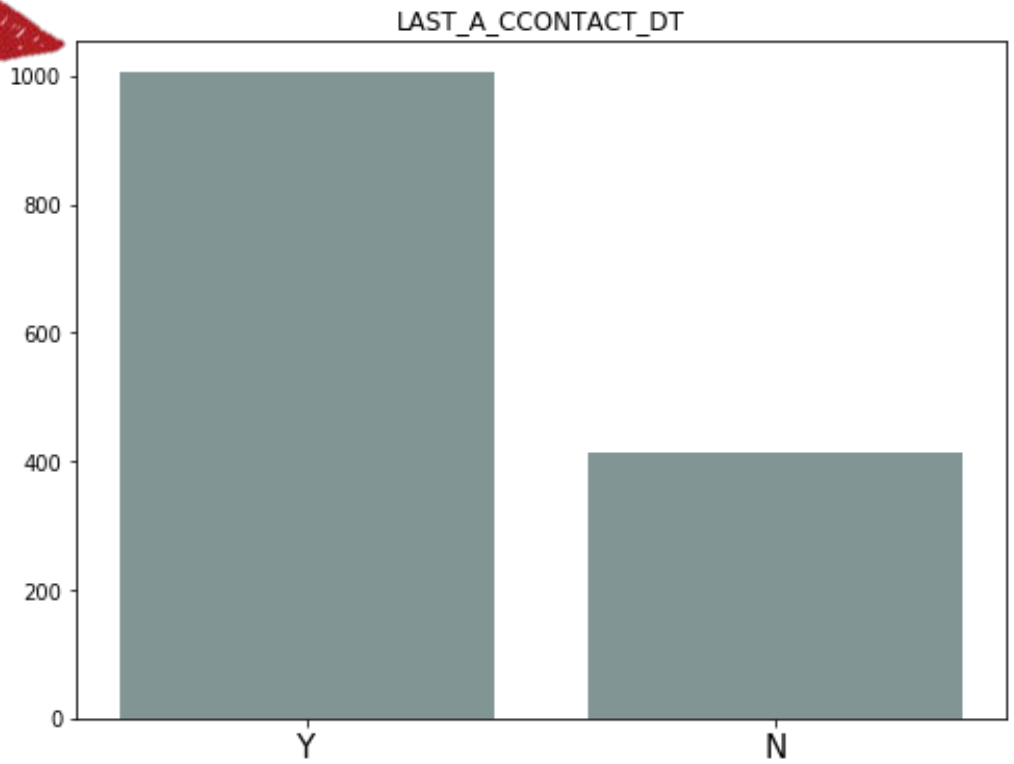
模型選擇與驗證

模型選擇與驗證 - 重要指標

影響「客戶是否會購買」的重要因素

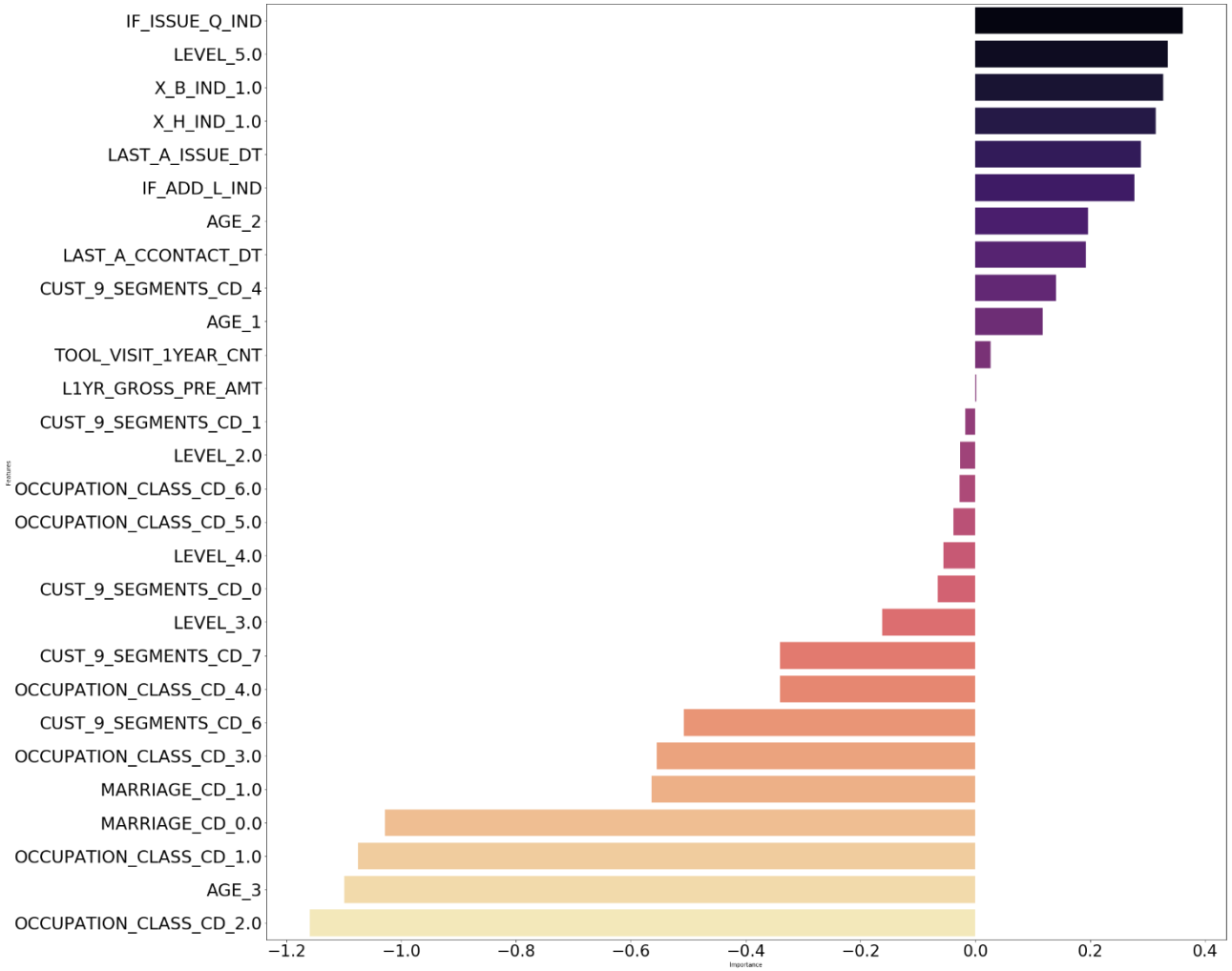


近三年是否有與 A 通路接觸



模型選擇與驗證 - 評估重要變數和購買的關係

Logistic regression係數長條圖



- 重要變數與「客戶是否會購買」的影響力
- 同一類別各情況影響購買程度高低

模型選擇與驗證

變數	變數的類別影響「客戶購買」的大小排序
年齡	中高>中>低>高
婚姻狀況	2>1>0
九大客群	E>C>B>A>H>G
客戶職業類別對核保風險程度	0>6>5>4>3>1>2

變數	變數的類別影響「客戶購買」的大小排序
往來關係等級	5>1>2>4>3
近一年業務員管理工具拜訪次數	Y>N
是否申辦 B 服務	
是否申辦 H 服務	
近三年是否有透過 A 通路投保新契約	
近三年是否有與A通路接觸	

PART 4

商業應用

商業應用- 效益說明

$$\text{Net Income(平均拜訪每個客戶的淨利)} = \frac{\text{TN} \times \text{B}_{\text{TN}} - \text{FP} \times \text{C}_{\text{FP}} - \text{FN} \times \text{C}_{\text{FN}}}{\text{N}}$$

Cost matrix

真實\預測	Y=0	Y=1
Y=0	C_{TP}=0	誤判購買客戶數 C_{FP}=1
Y=1	流失客戶數 C_{FN}=1	C_{TN}=0

Benefit Matrix

真實\預測	Y=0	Y=1
Y=0	B_{TP}=0	B_{FP}=0
Y=1	B_{FN}=0	B_{TN}=1

Y=0 : 不買
Y=1 : 買

B : Benefit
C : Cost

TP、TN、FP、FN：預測數目

商業應用- 模型比較

	Logistic Regression	Random Forest	Neural Network	XGBoost
AUC	0.7922	0.8041	0.7927	0.85195
Net Income	-0.29726	-0.29136	-0.2906	-0.2392

準確度高

無法反映真實的獲利方式

目標 考量真實情況，降低總成本

商業應用- 效益考量

- 分類正確 → 降低成本
- 對於流失的客戶有較高的損失

模型優化目標

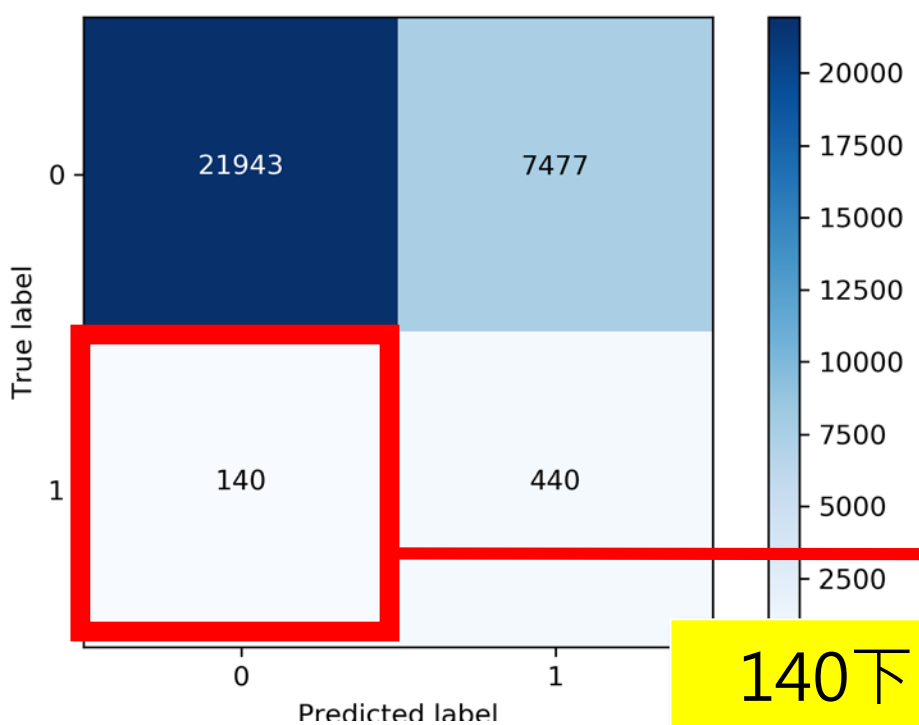
$$Cost\ Sensitive\ loss = \frac{1}{N}\sum C_{FN} \times FN + C_{FP} \times FP$$

真實\預測	不買 Y=0	買 Y=1
不買 Y=0	C_{TP}=0	C_{FP}=1
買 Y=1	C_{FN}=1	C_{TN}=0

C_{FN}=500

商業應用- 結果比較

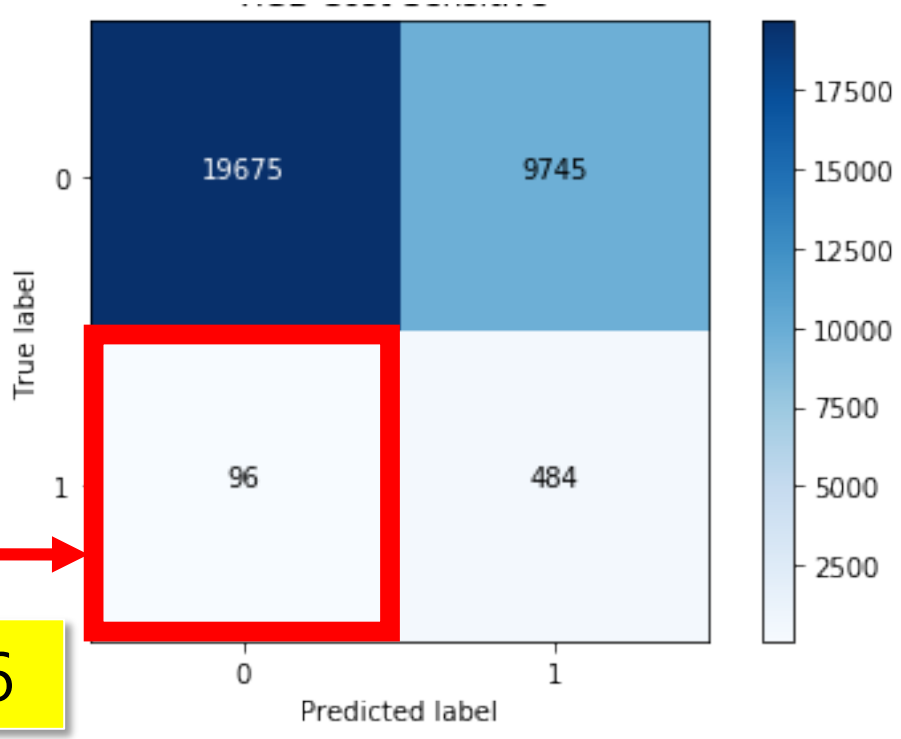
未考慮成本的預測結果



AUC	0.85195
Net Income	-0.2392

140下降至96

考慮成本的預測結果



AUC	0.824
Net Income	6.14

結論

模型分析結果考量保險公司實際的成本，能發揮最大效益

針對重點族群搭配合適的銷售手法，即可針對潛在客戶做深入推廣

THANKS