

南山人壽

理賠客戶再購與商品推薦

政大風管碩二 陳奕帆 政大風管四 何恬

台大財金所財工組碩一 周永昱 台大資工二 謝宗儒

大綱



研究問題 Research question



探索性資料分析 **EDA**



資料預處理 Data pre-processing



模型訓練 Model **Training**



附錄 **Appendix**



Research Question	EDA	Data pre-processing	Model Training	Appendix
		,		
				4
				4



探索性資料分析

EDA

- A. 理賠檔EDA
- B. 再購檔EDA
- C. 客戶屬性檔EDA

Research Question

EDA

Data pre-processing

Model Training

Appendix

A. 理賠檔EDA

- 共234428筆資料,13個feature
- 滿期金受益人RK有69%的Missing Value,生故保險金受益人RK有75%的Missing Value。此兩者應不適合做填補,但可用來產生更多feature,如:是否具滿期金受益人、是否具滿期金受益人、任一受益人是否為被保人...等。
- 理賠案件型態人數極度失衡,可以注意各類的再購率是否有明顯差異。尤其是當被保人死亡或 重病後,是否影響再購行為(此處須注意再購定義,如以被保人-被保人合併,那死亡件100%不 會有再購行為,可能要結合客戶關係檔,如被保人的一等親作為合併條件)
- 有97.96%的案件被保人等於事故人,其餘可能是家庭保單,因此取一位被保人當代表,而代表人並非事故人,因此產生被保人不等於事故人情況,因此理賠再購合併時應注意此種情況,避免漏掉再購。
- 事故人、要保人、被保人之間關係應仔細考慮,可搭配客戶關係檔做更多Feature Engineering。

Research Question	EDA	Data pre-processing	Model Training	Appendix
B. 再購檔EDA				
	資料,11個feature			
• 再購檔的資料	早是Dependent Variable	,可視所需來產生對應	的Y,例如是否再購、Ā	 再購什麼等。
• 產品細項欄位	1有多項只有一筆資料	l,是否該刪除此類資料	?	
• 大多數保單生	三效日在3~6月,是否	有什麼經濟意義?		

Research Question	EDA	Data pre-processing	Model Training	Appendix
C & S E IL INCOA				
C. 客戶屬性檔EDA				



資料預處理

Data pre-processing

- A. 理賠檔、再購檔、客戶屬性檔合併
- B. 合併檔案分析
- C. Deal with Miss Value
- **D.** Feature Engineering
- E. Categorical Variable Encoding
- F. Feature Scaling
- G. Deal with Imbalanced Data

9

Research Question EDA Data pre-processing Model Training Appendix

A. 理賠檔、再購檔、客戶屬性檔合併

- 1. 資料由要保人對要保人的方式進行合併
- 2. 並將理賠後120天內再購、180天內再購及360天內再購與否設為新變數,以便了解客戶再購情形
- 3. 本組也將合併後的資料作分析及解讀

)

Research Question	EI)A	Data pre-p	rocessing	Model Tra	nining	Appendi
B. 合併檔案分析							
不	同理賠案件	型態的再	購情形				
		55 由h	/+LL	400 点更赚	400点声畔		
	理賠案件型態	筆數	佔比	120內再購	180內再購	360內再購	
	身故給付	3841	1.64%	7.37%	7.86%	9.19%	
	完全失能	224	0.10%	7.59%	8.48%	9.82%	
	部分失能	160	0.07%	17.50%	19.38%	20.63%	
	重大疾病	6483	2.77%	3.44%	4.52%	7.11%	
	疾病醫療	132549	56.54%	4.30%	5.76%	9.67%	
	意外醫療	91171	38.89%	5.27%	7.22%	12.61%	
		234428	100%				

Research Question EDA Data pre-processing Model Training Appendix

B. 合併檔案分析

利用Scheffé法事後比較:

不同理賠案件型態的再購比例是否有顯著差異

- 不同理賠案件型態120天內的再購情形: 部分失能 > 完全失能 = 身故給付 > 意外醫療 > 疾病醫療 = 重大疾病
- 不同理賠案件型態在180天內的再購情形: 部分失能 > 完全失能 = 身故給付 = 意外醫療 > 疾病醫療 > 重大疾病
- 不同理賠案件型態在360天內的再購情形: 部分失能 = 意外醫療 > 完全失能 = 疾病醫療 = 身故給付 > 重大疾病

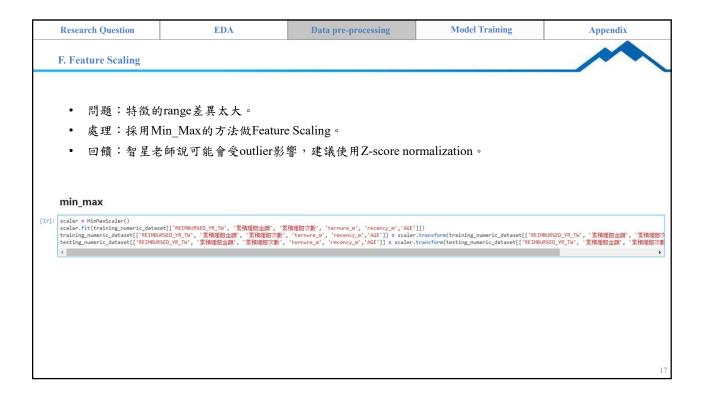
2

esearch Question	EDA		P	e-proces	8			el Train	-	Appendix
合併檔案分析										
:	理賠客戶中不同的	· 庆 庄 ź	活 Ul A	公 再 F	掛山く	刮吊:	甘雨日	生的	ロット	上面
•										C 15.1
		有再購比例						再購REG		
	01.傳染病和寄生蟲病	28.1%	25.0%	10.1%	26.1%	5.5%	4.2%	20.6%	8.4%	
	02.腫瘤	16.0%	15.3%	6.7%	19.3%	5.4%	8.6%	23.2%	21.5%	
	03.血液相關及免疫系統的疾患	23.5%	9.5%	5.4%	29.3%	12.8%	5.0%	25.2%		
	04.內分泌營養和代謝疾病 05.精神和行為疾患	16.8%	25.2%	4.1%	19.9%	2.2%	5.4%	34.1%		
		9.7%	15.9%	8.8%	14.8%	3.8%	4.9%	23.6%	28.0%	
	06.神經系統疾病	14.4%	24.6%	4.0%	17.1%	6.9%	10.9%	24.0%	12.6%	
	07.眼和附器疾病 08.耳和乳突疾病	17.6%	13.7%	5.0%	21.4%	8.4%	8.6%	26.5%		
	08. 年刊孔关疾病	23.0%	18.2%	10.3%	27.1%	7.0%	6.1%	22.9%	8.4%	
		15.7%	13.8%	5.5%	24.2%	8.1%	7.6%	23.5%		
	10.呼吸系統疾病 11.消化系統疾病	25.5%	26.4%	10.6%	25.1%	5.5%	3.7%	19.8%		
	11.岁10系然疾病 12.皮膚和皮下組織疾病	20.6%	17.0%	5.3%	22.5%	7.7%	8.9%	25.5%	13.1%	
	12.皮膚和皮下組織疾病 13.肌肉骨骼系統和結締組織疾	20.7%	23.0%	6.6%	28.1%	5.0%	5.0%	22.0%	10.4%	
	13.肌肉膏脂系統和結締組織疾		16.3%	4.4%	24.2%	6.5% 8.4%	7.0%	26.1%	15.5%	
	14. 心脉 土 组 系 航 疾 病 15. 妊娠、 分娩 和 產 褥 期	19.8%	20.8%	8.3%	22.1%		6.3%	23.3%	10.8%	
	15.妊娠、万烷和產份期 17.先天畸形變形和染色體異常	35.1% 23.9%	28.3%	18.9%	26.4% 15.6%	3.0% 9.4%	3.1% 0.0%	14.1% 25.0%		
	18.症狀異常所見,不可歸類		34.4%	6.3% 9.8%	25.0%	5.4%				
	19.損傷中毒和外因的某些其他	21.7%	17.8%				5.3%	23.3%	13.4%	
	20.疾病和死亡的外因		13.1%	4.8%	45.2% 28.1%	2.4% 7.8%	1.2% 6.2%	22.6% 25.7%	10.7% 10.1%	
	20.疾病和光亡的外凸 21.影響健康狀態與保健機構接	27.6%	16.0%	6.1%						
	99.不知道不想猜		17.4%	9.0%	33.3%	5.8%	4.3%	22.8%		
	99. 个知道个忠侗	27.8%	16.4%	7.0%	27.9%	6.9%	6.2%	25.7%	10.0%	

Research Question	EDA	Data	pre-processing	g	Model Training		Appendix
C. Deal with Missing V	alue						/
 刪除且Missin 	g Value的Feature刪除	, 如年的	女入、婚姻	1狀況和約	魚資產筆。		
	·欄位為空值的列刪除		~	1))CDG1 - W	0只在 1		
	1000 1 - 1000 1 - 1000 1						
		E	BundleSubtype2	illness_code	DiagnosisCode_DESC	claim_settle_dt	REIMBURSED_YR_
		0	5.N疾病醫療	C18	02.腫瘤	2015-03-25	46987
		1	5.N疾病醫療	C18	02.腫瘤	2015-08-15	7087
		2	5.N疾病醫療	C18	02.腫瘤	2015-08-15	30712
		3	5.N疾病醫療	C18	02.腫瘤	2015-08-15	7087
		4	5.N疾病醫療	C18	02.腫瘤	2015-08-15	7087
			•••				
		210784	6.N意外醫療	Y99.8	99.不知道不想猜	2017-10-15	1050
		210785	6.N意外醫療	V23	20.疾病和死亡的外因	2017-02-19	813
		210786	6.N意外醫療	Y99.8	99.不知道不想猜	2017-08-23	1627
東攵	理後匯入資料如右圖 →	210787	5.N疾病醫療	D36	02.腫瘤	2017-08-11	546
至	生皮些八貝附如口圖 フ	210788	4.C重大疾病	125.1	09.循環系統疾病	2017-04-29	16825

Resear	ch Question	EDA	Data pre-processing	Model Training	Appendix
D. Fea	ture Engineerin	g			
1.	累計理賠金	額			
2.	現有/曾有	AH保單種類數			
3.	要被保人相	同與否			
4.	要保人與受	益人相同與否			
5.	疾病發生部	位			

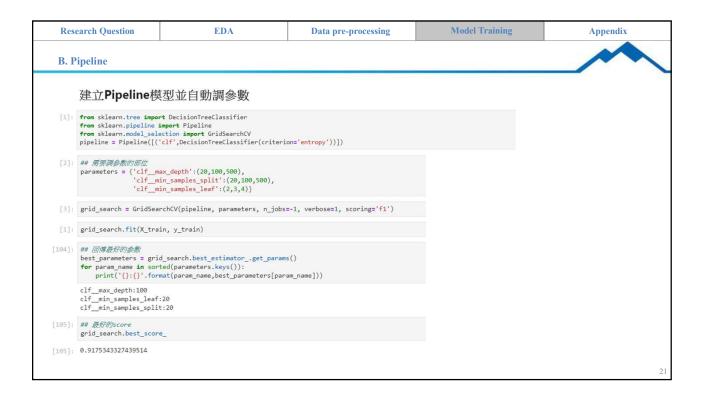






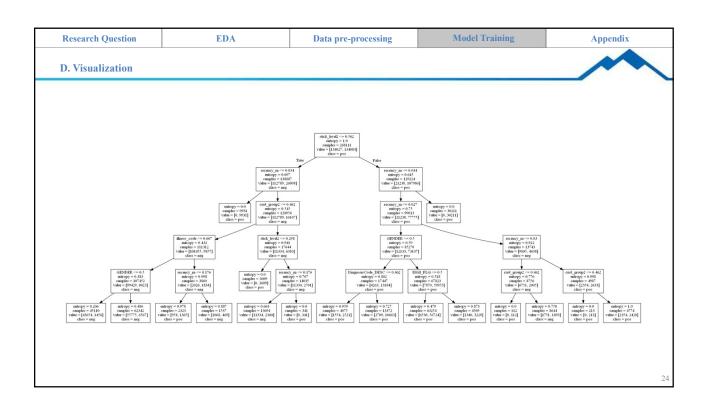


Research Question	EDA	Data pre-processing	Model Training	Appendix
A. Baseline				
				20



Research Question		EDA		Data pre-processing	Model Training	Appendix
C. Evaluation						
混淆矩	車解讀					
Expression (11.1	A constitution of the	otoriesto escato	1 - 16 - 17		
nttps://www.	ibinx.com/2018	/understandii	ng-skieari	n-classification-report/		
[96]: from sklear	n.metrics impo	rt classifi	cation_r	eport		
	= grid_search					
print(class	ification_repo	rt(y_test,	predicti	ons))		
	precision	recall f	1-score	support		
	0.92	0.92	0.92	66056		
	0.92	0.92	0.92	65999		
accurac	/		0.92	132055		
macro av	g 0.92	0.92	0.92	132055		
weighted av	0.92	0.92	0.92	132055		
107]: print('Trai	ning data scor	re: {}'.form	at(grid	search.score(X_train,	v train)))	
				search.score(X_test,		
Training da	ta score: 0.93	62802709885	092			
Training da	ta score: 0.92	00461914819	26			







B. 備註

C. EDA result

D. Reference

- 2

Research Question EDA Data pre-processing Model Training Appendix

A. 分工

- 周永昱: EDA、資料預處理、機器學習、簡報製作
- 謝宗儒:學習了機器學習相關:KNN、回歸演算法、決策樹、隨機森林、降維 演算法、貝葉斯演算法、編碼方式;保險知識、資料前處理:醫療保險、意外 險、壽險等保單種類跟概況
- 何恬:理賠再購資料合併分析、資料特徵解讀及選擇、新增延伸特徵、特徵類型轉換、簡報製作
- 陳奕帆 : 客戶屬性變數分析、合併檔資料分析、再購情形事後比較、不同疾病的再購比例分析、尋找可增加特徵

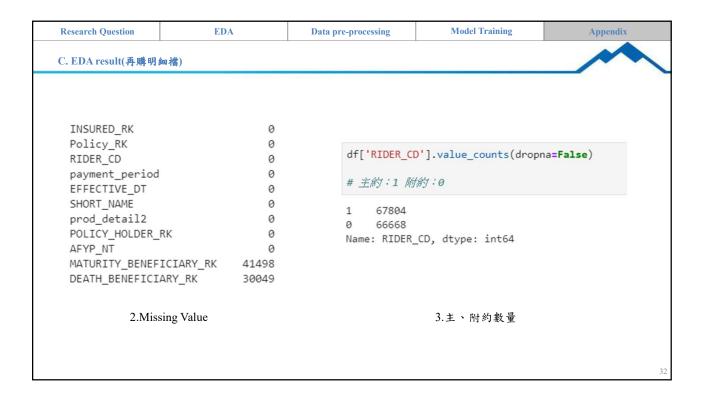


Research Question	EDA	Data pre-processing	Model Training	Appendix
C. EDA result(理賠檔)				
(= ::= 14)				
_		理賠檔(CLAIM_ACC	T FINIU期价的明	-
	欄位	型品值(CLAIM_ACC 類型	- 1_FIIV) 懒 型	
	I INJURED_RK		事故人RK	
	Claim_RK	字元	SHIP BURELUNG COMMUNICATION	1
	Policy_RK	字元	保單號碼	7
4	BundleSubtype2		理賠案件型態	
	illness_code	字元	疾病代碼	
	illness_desc		疾病名稱	
	DiagnosisCode_DES		疾病分類名稱	
	claim_settle_dt	日期	理賠結案日期	
	REIMBURSED_YR_		理賠金額(歸至該結案年度)	
1	0 INSURED_RK	字元	被保人RK	
1	1 POLICY_HOLDER_F		要保人RK	
	2 MATURITY_BENEFI		滿期金受益人RK	
1	3 DEATH_BENEFICIA	RY_RK 字元	生故保險金受益人RK	
		1 lgg (>+ - ntl</td <td></td> <td></td>		
		1. 欄位說明		



Research Question	EDA	Data pre-processing	Model Training	Appendix
C. EDA result(理賠檔)				
C. EDA result(理赔檔) W18	26959 8 22910 15598 9080 7 8174 0 1 1	事故人被保人要保. [15]: # 專您(=發係(-要係)	DER_RK']) & (df['INJURED_RK']==df['POLIC DER_RK'])]['Policy_RK'].count() DER_RK'])]['Policy_RK'].count()	Y_HOLDER_RK'])]['Policy_RK'].count()
4. 4	各項疾病人數	5. 事故	人、被保人、要保人重複	复情况
				30

EDA result(再	-購明細檔)				
		再購明	細檔(COV_ACCT_FIN)相	順位說明	
	欄位	類型	名稱	說明	
1	INSURED_RK	字元	被保人RK		
2	Policy_RK	字元	保單號碼		
3	RRKER_CD	數值	主附約註記	1=主約,0=附約	
4	payment_period	字元	保費繳法		
5	EFFECTIVE_DT	數值	保單生效日		
6	SHORT_NAME	字元	產品細項		
	prod_detail2	字元	產品類型	REG=期繳商品,SIN=躉繳商品,ILP=投資型商品,	
7	NO 5+0H3			AHa=住院手術商品,AHb=重疾癌症,,AHc=意外	
				保障,AHd=長期照顧	
8	POLICY_HOLDER_RK	字元	要保人RK		1
9	AFYP_NT	數值	保單保費		1
10	MATURITY_BENEFICIARY_RK	字元	滿期金受益人RK		
11	DEATH_BENEFICIARY_RK	字元	生故保險金受益人RK		
	-			-	
			* [HH () \ () HH		
			1.欄位說明		





Research Question	EDA	Data pre-processing	Model Training	Appendix
C. Reference				