



富邦人壽業務員 績效數位優化方案

第二組

倪采靖、翁梓鴻、郭家亨、陳采妍、謝明穎

目錄 <<<

01

背景

問題及目標

02

資料觀察

監督式學習分析

03

機器學習模型分析

Linear Regression, Decision Tree, Ensembling,
SVM, LSTM

04

結論

持續優化之處



01

背景

問題

- FBFLI客戶定聯系統
- 員工對系統的依賴性不強
- 透過數據分析與數位工具協助業務員提升績效



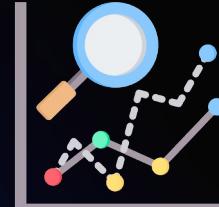
目標



數據分析



找出潛在因素



使用模型預測



提供富邦
優化內容



02

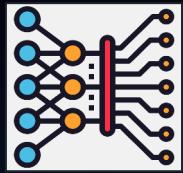
資料觀察

使用監督式學習分析，製圖觀察業務員基本資料與績效之關係

數據



Y值使用保費的
平均值

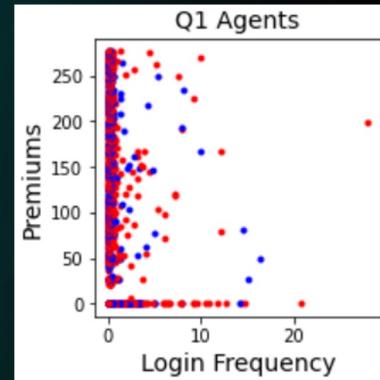
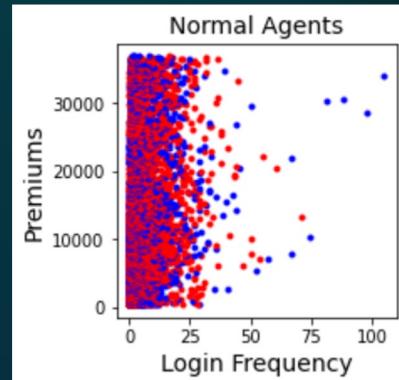
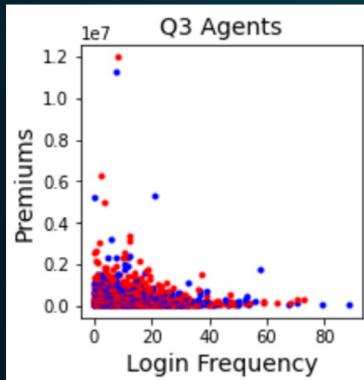


將保費收入分成三層：
1. > 上四分位數
2. < 下四分位數
3. 下四分位數 ~ 上四
分位數

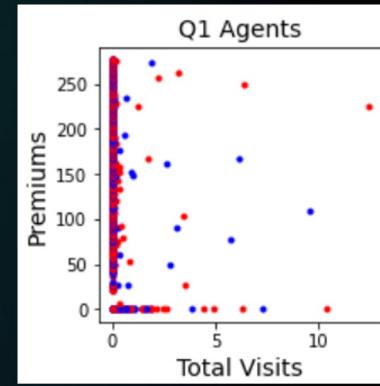
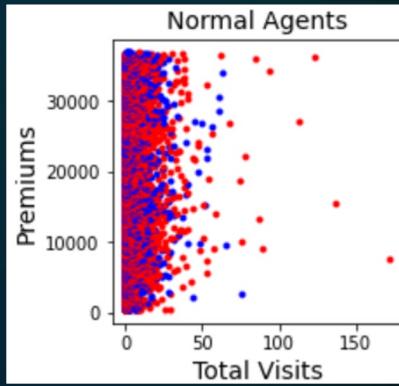
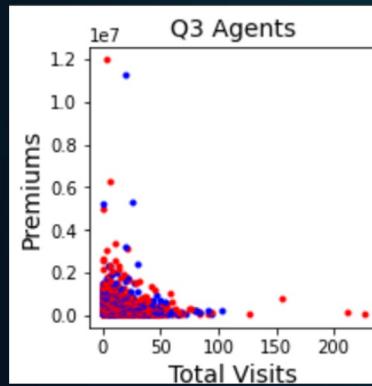


繪製圖表，觀察各項
業務員基本資料與業
績的關係

登入次數：

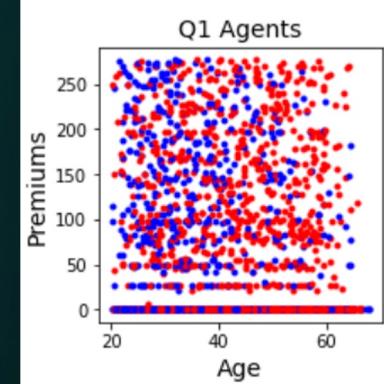
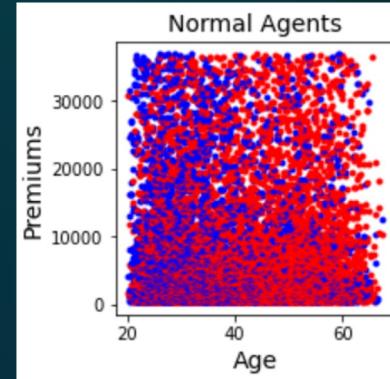
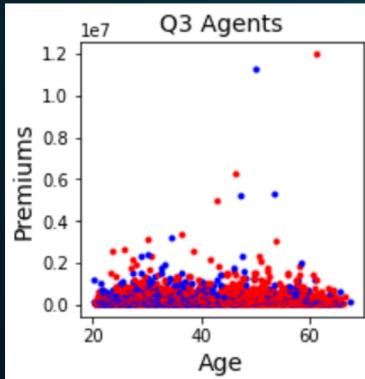


拜訪客戶次數：

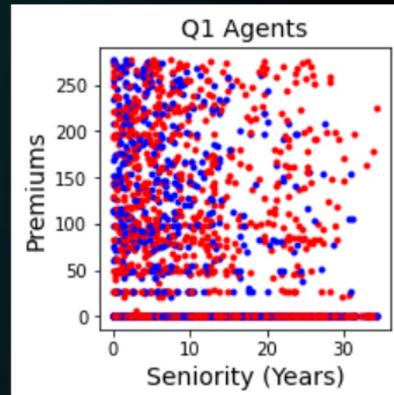
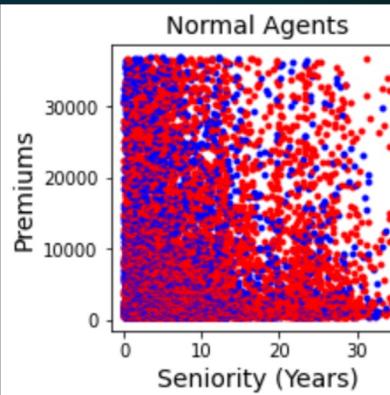
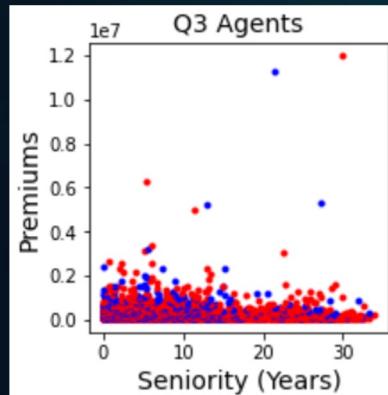


Male : Blue, Female : Red

年齡：

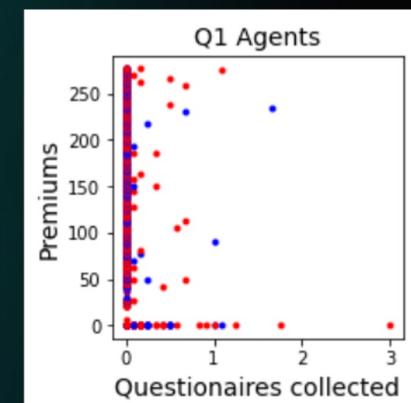
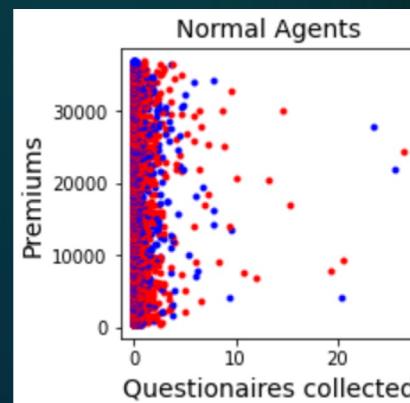
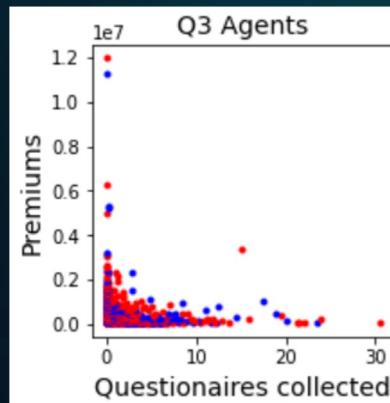


年資：

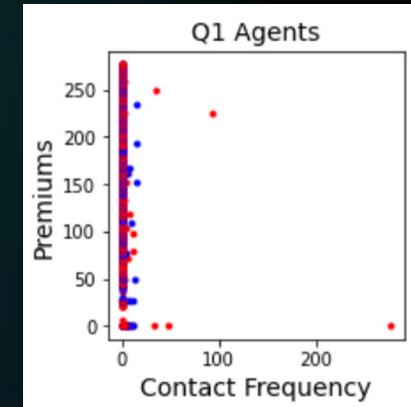
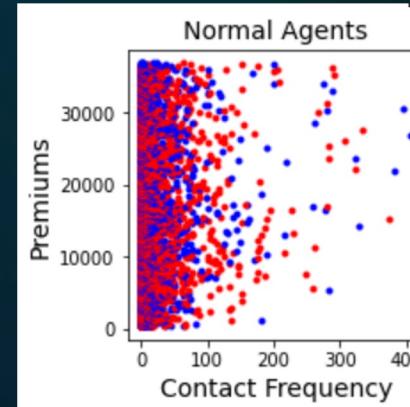
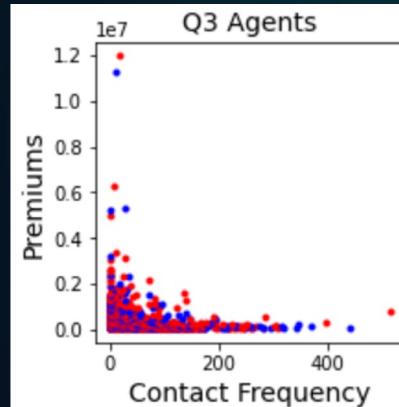


Male : Blue, Female : Red

問卷填達數 :



定聯次數 :



Male : Blue, Female : Red



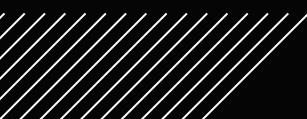
03



機器學習模型分析



Linear Regression, Decision tree, Ensembling, SVM,
LSTM



03-1

Linear Regression

One_hot encoding

→ 線性迴歸之結果無法有效預測數值

```
. reg sumfyc agent_title_1 agent_title_2 agent_title_3 agent_title_4 agent_title_5 agent_title_6 agent_title_7 agent_title_8 agent_t  
> itle_9 agent_title_10 agent_title_11 agent_title_12 agent_title_13 agent_title_14 agent_title_15 agent_title_16 agent_age on_board  
> _age seniority agent_sex city_1 city_2 city_3 city_4 city_5 city_6 city_7 city_8 city_9 city_10 city_11 city_12 city_13 city_14 ci  
> ty_15 city_16 city_17 city_18 city_19 city_20 customer_cnt memo_length login_cnt contact_cnt contact_cust_visit_cnt customer_v  
> isit_cnt qnr_cnt consult_cnt consult_cust_l_profyc l_fyc l_propie l_pie l_pre l_propre  
note: agent_title_11 omitted because of collinearity  
note: city_19 omitted because of collinearity
```

Source	SS	df	MS	Number of obs	=	53,592
Model	1.1477e+13	54	2.1253e+11	F(54, 53537)	=	271.82
Residual	4.1860e+13	53,537	781881690	Prob > F	=	0.0000
				R-squared	=	0.2152
				Adj R-squared	=	0.2144
Total	5.3336e+13	53,591	995247198	Root MSE	=	27962

通過箱線圖直觀的顯示數據分布，並觀測數據中的異常值。箱線圖一般由五個統計值組成：最大值、上四分位、中位數、下四分位和最小值。一般來說，觀測到的數據大於最大估計值或者小於最小估計值則判斷為異常值，其中

最大估計值=上四分位+1.5*(上四分位-下四分位)
最小估計值=下四分位-1.5*(上四分位-下四分位)



比上邊緣高的值就是異常值

雖已經處理Y(產壽險佣金總和)之極值

→ 線性迴歸之結果無法有效預測數值

▼ 2.5.5 模型测试

加载训练好的模型参数，在测试集上得到模型的MSE指标。

```
[ ] # 加载模型权重  
runner.load_model(saved_dir)  
  
mse = runner.evaluate(test_dataset)  
print('MSE:', mse.item())  
  
MSE: 986801664.0
```

▼ 2.5.6 模型预测

使用 Runner 中 `load_model` 函数加载保存好的模型，使用 `predict` 进行模型预测，代码实现如下：

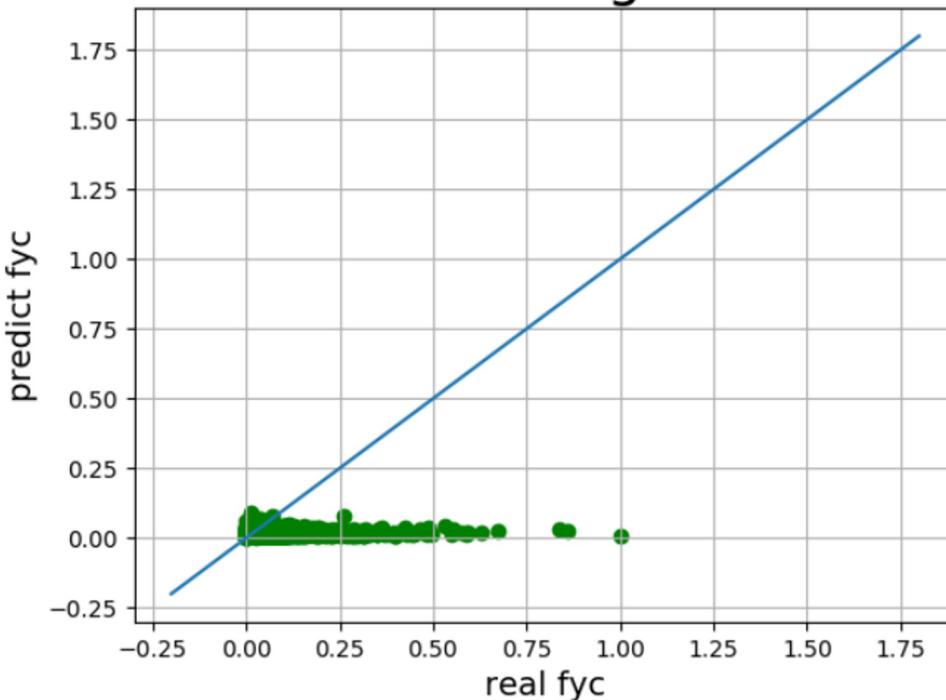
```
[ ] runner.load_model(saved_dir)  
pred = runner.predict(X_test[3:4])  
print("真实FYC:", y_test[3:4].item())  
print("预测的FYC:", pred.item())
```

```
真实FYC: 16483.0  
预测的FYC: 17935.86328125
```

maxminscaler:

→ 線性迴歸之結果無法有效預測數值

fubang

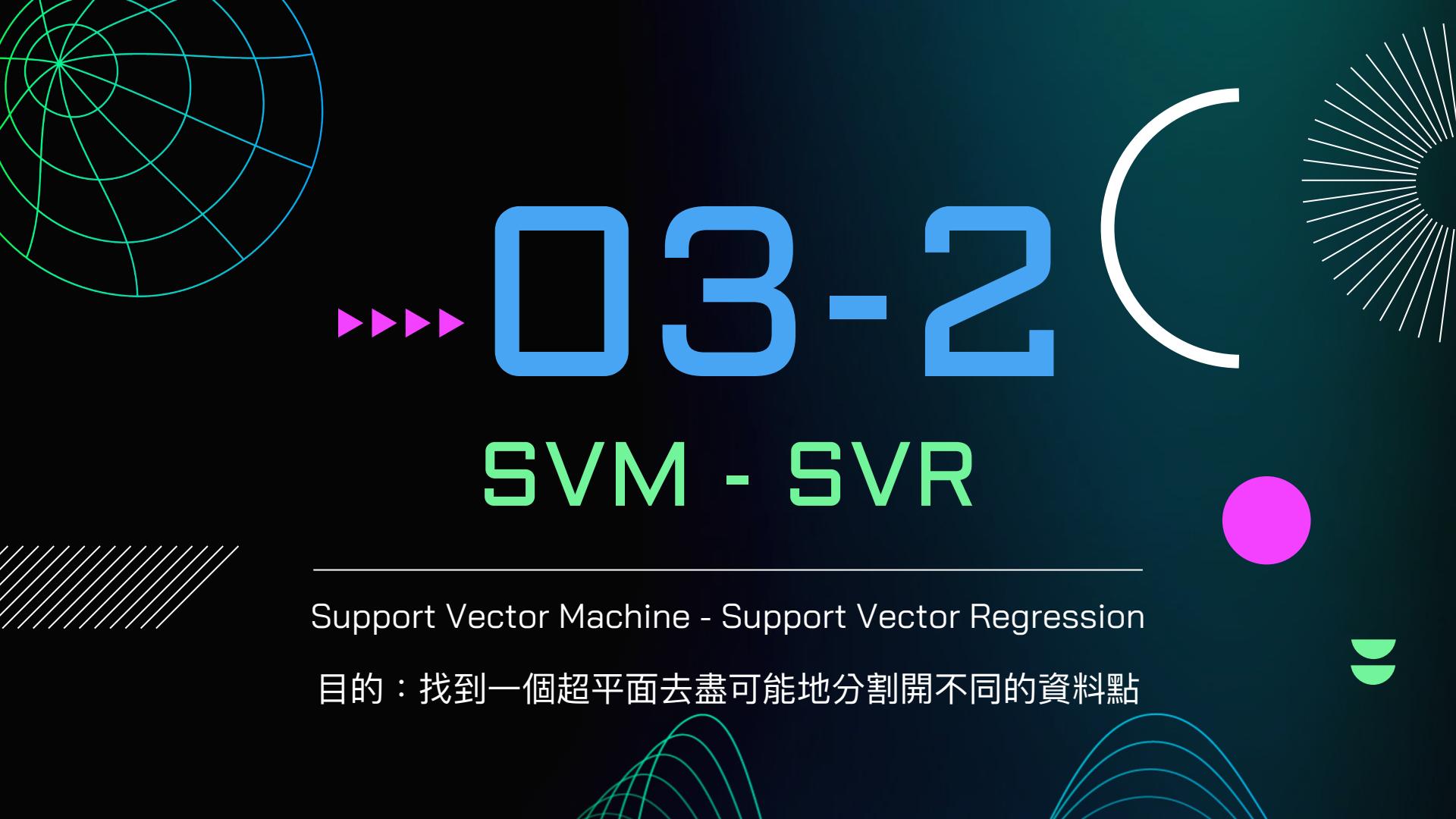


score of lr: -0.03234725621615597

```
[ ] y_tests[50:55]
array([[0.04597612],
       [0.0258264 ],
       [0.00746568],
       [0.03160405],
       [0.09794771]])
```



```
[ ] lr_y_predict[50:55]
array([[0.01281698],
       [0.01274899],
       [0.01727256],
       [0.02771023],
       [0.01609619]])
```



03-2

SVM - SVR

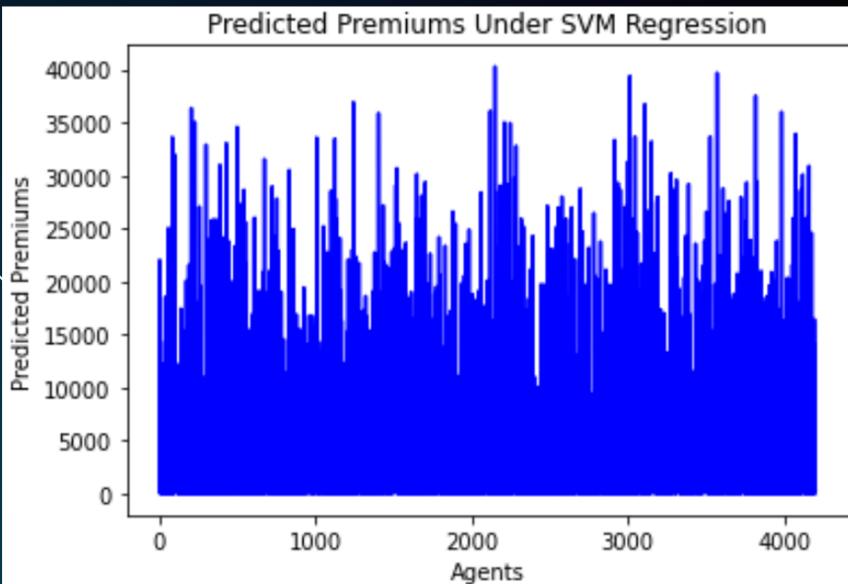
Support Vector Machine - Support Vector Regression

目的：找到一個超平面去盡可能地分割開不同的資料點

使用資料

- 訓練集用所有資料（包含0值和極端值）
 - 不捨棄0值：有可能是新進業務員，很認真在做很多動作，但就是還沒有創造保費收入
 - 不捨棄極端值：彌補0值的缺失
- 測試集用隨機的1/4組資料

分析結果



	real	predict	error
0	634.666667	151.526063	483.140604
1	27727.041667	22034.898024	5692.143643
2	0.000000	1.769352	-1.769352
3	337.083333	0.377295	336.706038
4	8839.458333	3084.117093	5755.341240
5	5687.125000	3889.553496	1797.571504
6	373.416667	197.667327	175.749339
7	430282.625000	14375.735366	415906.889634
8	54.833333	29.828734	25.004600
9	1730.500000	52.907257	1677.592743
10	45067.083333	3305.258827	41761.824506
11	2062.166667	1495.853588	566.313079
12	671.833333	47.721895	624.111439
13	3258.500000	466.335144	2792.164856
14	480515.958333	185.379828	480330.578506
15	0.000000	26.663503	-26.663503
16	1020.750000	148.049974	872.700026
17	6224.708333	7343.441277	-1118.732944
18	560.708333	516.383615	44.324718
19	87931.250000	8368.130617	79563.119383

Mean error = 32774.415195408

Absolute mean error = 33691.529029052326

Maximum error = 2665630.9049273445

Minimum error = -32477.55550045637

Mean (predict values) = 4351.27119949368

Mean (y_mean) = 37125.686394901684

列舉前20筆資料去做參考

03-3

Decision Tree

參數設定

- Criterion
 - Square error, friedman mse, absolute error, poisson
- Splitter
 - Best
- Min Sample Leaf
 - 2~20
- Best Model - MAE 84959

```
model = DecisionTreeRegressor(criterion='squared_error',
| min_samples_leaf=19, splitter='best', random_state=5)
```

其他前處理&交叉驗證

- 類別變數
 - Ordinal, One hot, Drop
- 缺失值
 - 0, Bfill, Ffill, mean
- 極端值
 - Isolation Forest, Std
- 日期
 - 轉成自1950年開始的月份數
- 交叉驗證00000



3-4

Ensembling



Random Forest

- Criterion
 - Square error, friedman mse, absolute error, poisson
- N Estimators
 - 10~50
- Min Leaf Nodes
 - 2~20
- Best Model - MAE 78473

```
rnd_clf = RandomForestRegressor(n_estimators=10, max_leaf_nodes=20,  
| n_jobs=-1, random_state=42, criterion='poisson')
```

Gradient Boosting

- Criterion
 - Square error
- N Estimators
 - 2~10
- Max Depth
 - 2~30
- N Iter No Change
 - 10
- Learning Rate
 - 0.1~0.01
- MAE 58962

```
gbrt_best = GradientBoostingRegressor(  
    learning_rate=2.1040543606193775e-10, max_depth=29,  
    n_estimators=9, random_state=42, n_iter_no_change=10)
```

Hist Gradient Boosting

- 預設值
- MAE 83971

```
hgb_reg = HistGradientBoostingRegressor(random_state=42)
```

Stacking

- Estimators
 - Random forest, Linear SVR, Gradient boosting, Hist Gradient Boosting
- Final Estimators
 - Gradient Boosting
- MAE 68732





3-5

LSTM

非線性特徵的 Logistic Regression

激勵函數: ReLU

時間序列

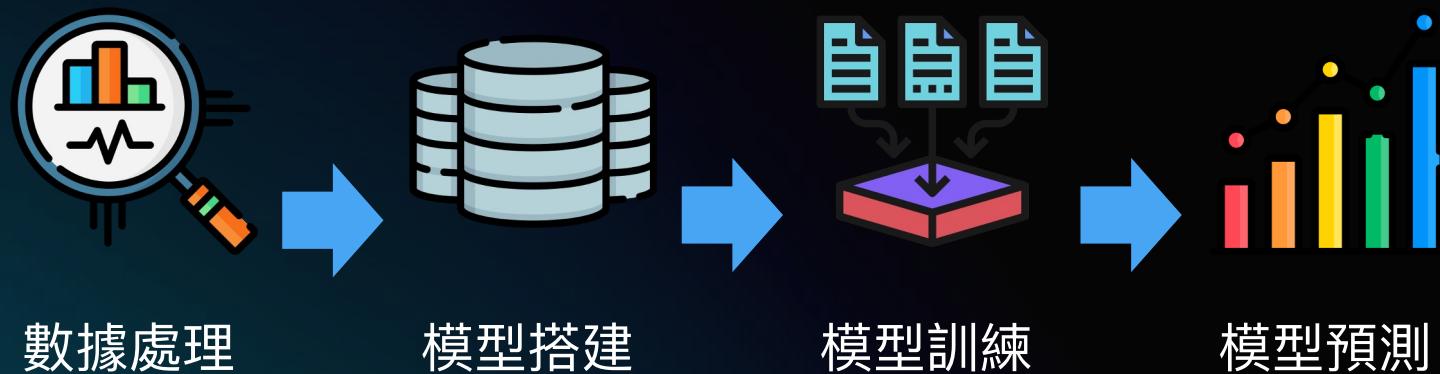
AGENT_ID
AGENT_AGE
ON_BOARD_AGE
SENIORITY
AGENT_SEX
CUSTOMER_CNT
MEMO_LENGTH
LOGIN_CNT
CONTACT_CNT
VISIT_CNT
QNR_CNT
CONSULT_CNT
I_PROPERTY_INSURANCE_FYC
I_FYC
sumfyc

Base Line

AGENT_ID
AGENT_AGE
ON_BOARD_AGE
SENIORITY
AGENT_SEX
CUSTOMER_CNT
MEMO_LENGTH
LOGIN_CNT
CONTACT_CNT
CONTACT_CUST_CNT
VISIT_CNT
CUSTOMER_VISIT_CNT
QNR_CNT
CONSULT_CNT
sumfyc

使用 ID < 35,000 的資料做為訓練集，其餘做為測試集

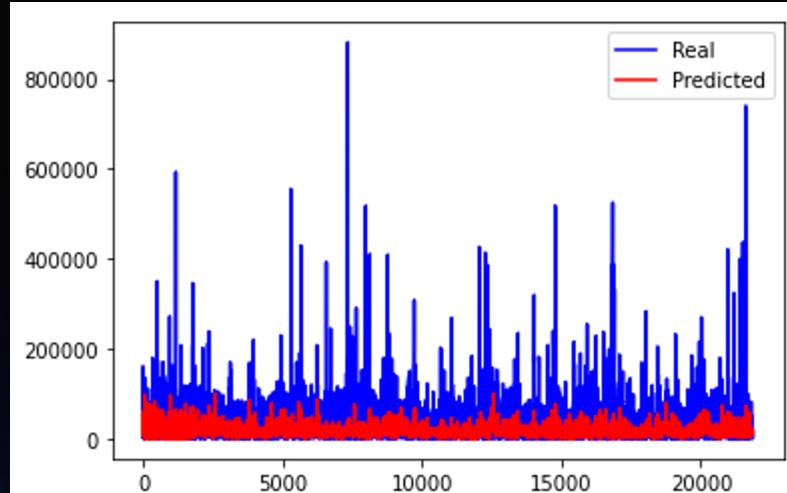
分析方式



預測結果

Base line

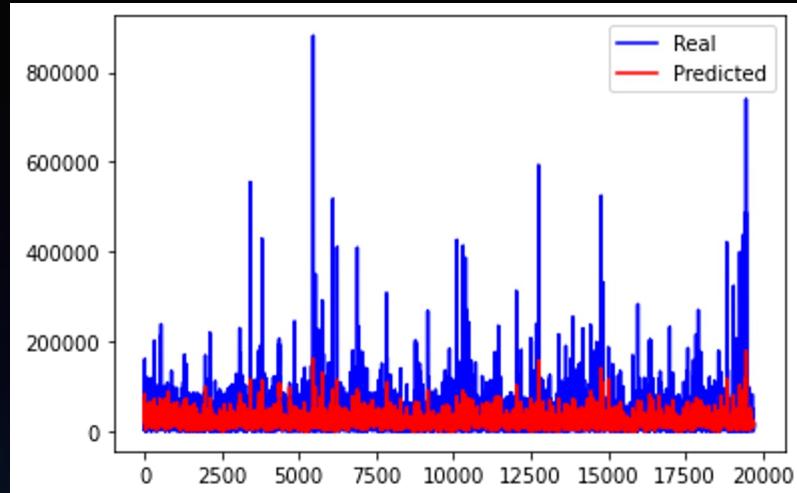
```
[ ] 100 - mean_absolute_error(y_true, y_pred)  
-15649.778258040527
```



預測結果

時間序列

```
100 - mean_absolute_error(y_true, y_pred)  
-14950.984180905542
```



One hot encoding

困難點

記憶體不足

```
In [2] train_df.shape, test_df.shape  
((53592, 130), (12573, 130))
```

模型训练

```
In [9] EPOCH_NUM = 100 # 设置外层循环次数  
TRAIN_BATCH_SIZE = 1000 # 设置batch大小  
BATCH_SIZE = 1000 # 设置batch大小  
training_data = train_df.iloc[: -10000].values.astype(np.float32)  
val_data = train_df.iloc[-10000: ].values.astype(np.float32)
```

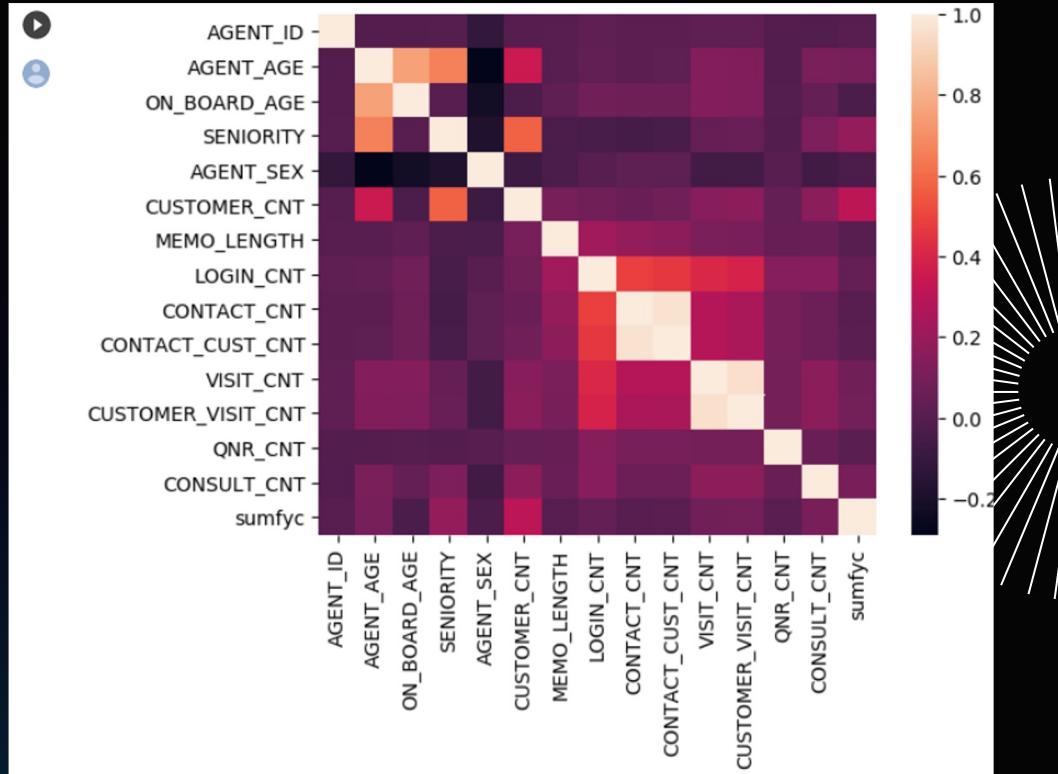
困難點

取有效的x

圖中最和sumfyc有相關性：

年資

客戶數量



困難點

職級:

主任級才會有正相關
(業務主任 行銷主任)

R-squared

= 0.0751

```
. . . reg sumfyc agent_title_1 agent_title_2 agent_title_3 agent_title_4 agent_title_5 a  
> _title_9 agent_title_10 agent_title_11 agent_title_12 agent_title_13 agent_title_14
```

Source	SS	df	MS	Number of obs	=	53,592
Model	4.0078e+12	15	2.6719e+11	F(15, 53576)	=	290.20
Residual	4.9328e+13	53,576	920719269	Prob > F	=	0.0000
Total	5.3336e+13	53,591	995247198	R-squared	=	0.0751
				Adj R-squared	=	0.0749
				Root MSE	=	30343

sumfyc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
agent_title_1	-45099.82	9150.574	-4.93	0.000	-63035.02 -27164.62
agent_title_2	-21227.94	9162.657	-2.32	0.021	-39186.82 -3269.057
agent_title_3	-30251.98	9159.851	-3.30	0.001	-48205.37 -12298.6
agent_title_4	-34100.22	9153.202	-3.73	0.000	-52040.57 -16159.87
agent_title_5	-29212.6	9211.706	-3.17	0.002	-47267.62 -11157.58
agent_title_6	-31964.25	9180.799	-3.48	0.000	-49958.69 -13969.81
agent_title_7	-25704.44	9198.503	-2.79	0.005	-43733.58 -7675.298
agent_title_8	-21865.71	9438.542	-2.32	0.021	-40365.33 -3366.093
agent_title_9	-2885.459	9250.37	-0.31	0.755	-21016.26 15245.34
agent_title_10	-153.3637	12938.45	-0.01	0.991	-25512.83 25206.11
agent_title_11	15717.96	13638.32	1.15	0.249	-11013.27 42449.19
agent_title_12	-57328.95	10649.06	-5.38	0.000	-78201.18 -36456.71
agent_title_13	-35538.08	9414.106	-3.77	0.000	-53989.81 -17086.36
agent_title_14	40792.07	13638.32	2.99	0.003	14060.84 67523.3
agent_title_15	-48082.82	13257.97	-3.63	0.000	-74068.54 -22097.09
_cons	60246.82	9148.866	6.59	0.000	42314.96 78178.67

除了分析產品對公司的貢獻之外，我們也使用分群演算法來將每個業務員每年的貢獻進行分類，不同組的業務員的貢獻有不小的差距，因此，我們觀察到業務員在每年的表現中是有存在不同的差異。

利用上述的方法對公司有基本的了解之後，我們分析長期在公司工作的業務員，可以透過這些業務員來了解他們長期的發展跟公司之間的關係，經過分群演算法的計算後，大致上可以將這些業務員分成兩類，貢獻度較高的業務員人數大概佔所有業務員人數的百分之二十，而這些業務員的業績大概佔整間公司的業績之百分之八十，當我們將這相同的模型套用到每一年所有的業務員上，我們找到了相同的結論。

除了有八二二十法則存在於這間公司，我們也觀察到這間公司的產品與獎勵制度間的關聯非常重要，如果公司發行的產品符合市場需求並且有提出吸引業務員的獎勵制度，他們的收入可以在短時間內提高許多，如果在未來公司能夠符合上述的兩個條件，可以幫助他們的業績在短時間內迅速成長。

論文名稱: 壽險業務員績效與業績目標關係之分析研究

論文名稱(外文): On the Relationship of Insurance Agents Performance and Sales Target

指導教授: 曹承璉、吳玲玲

指導教授(外文): Seng-Cho Chou、Ling-Ling Wu

口試委員: 周子元

口試委員(外文): Tzy-Yuan Chou

口試日期: 2018-07-17

學位類別: 碩士

校院名稱: 國立臺灣大學

系所名稱: 資訊管理學研究所

壽險業務員績效與業績目標關係之分析研究

困難點

冗餘信息、過擬合(收斂太快)

特徵工程

Ex. 提取特殊的部份(細分為季、月份...)作为新的屬性

交叉特徵 (ex. 職級*城市)

閾值的選取: ex. 大致上可以將這些業務員分成兩類，貢獻度較高的業務員人數大概佔所有業務員人數的百分之二十，而這些業務員的業績大概佔整間公司的業績之百分之八十

困難點

Maxminscaler 與否
(會大量影響權重)

困難點

其他模型:梯度增強

LightGBM，XGBoost，CatBoost

時間序列適合用樹模型

通過構造一組弱的學習器（樹），並把多顆決策樹的結果累加起來作為最終的預測輸出。

模型的融合

困難點

調整參數(ex. Learning_rate, layer 數量)

可能可以改進結果

04

結論

持續優化之處

公司面



預測優秀業務員較容易達成，
激勵員工成為優秀業務員較困難。
目前的數位活動對業績刺激不明顯，
可能需要再進一步：例：對客戶分群、為客戶分配適合業務員
若對員工特徵能有更多了解 (ex.性格) 分群可以更加詳盡。

員工面



數位活動的使用與推廣、長期客戶數積累。
超級業務員：賣壽險月入百萬





感謝聆聽！

