

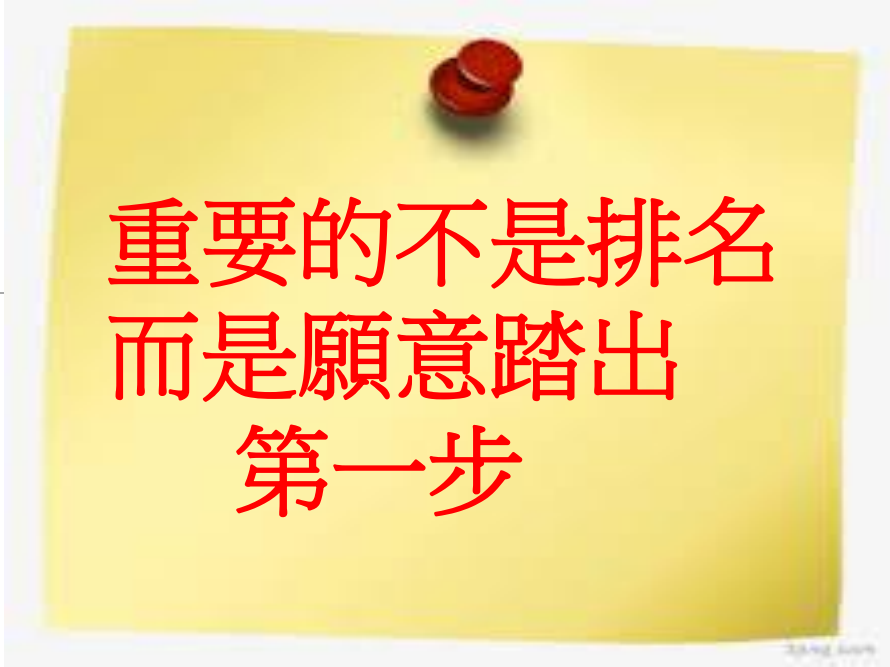
# T-Brain AI實戰吧

## -客戶續約金額預測實戰

### 心得分享

主講人：薛百惠

2019/03/29



重要的不是排名  
而是願意踏出  
第一步

# Outline

---

- 競賽主題說明
- 競賽資料說明
- 心得分享

# 競賽主題說明

➤ 跨國產險公司提供近一年的客戶特徵與續約金額狀況，找出有效預測既有客戶的續約金額模型或方法

- ✓ 開發新客戶所需要的成本是維護既有客戶的5倍
- ✓ 掌握影響客戶續約或流失，為企業經營的重要課題
- ✓ <https://tbrain.trendmicro.com.tw/Competitions/Details/3>



495  
參賽隊伍



總獎金  
新台幣 20 萬元

開始 7/23/2018

結束 9/14/2018

# 競賽主題說明

## ➤ 評分方式

### ✓ MAE (Mean Absolute Error) 平均絕對誤差

在統計中，平均絕對誤差（MAE）是衡量兩個連續變量之間的差異的指標。假設X和Y是表示相同現象的配對觀察值的變量。Y與X的例子包括預測值與觀測值的比較，隨後的時間與初始時間的比較，以及一種測量技術與另一種測量技術的比較。考慮n個點的散點圖，其中點i具有坐標（ $x_i$ ， $y_i$ ）...平均絕對誤差（MAE）是每個點與 $Y = X$ 線之間的平均垂直距離，其也被稱為一對一線。MAE也是每個點與 $Y = X$ 線之間的平均水平距離。

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}$$

# 競賽主題說明

## ➤ 競賽規則

- ✓ 得獎隊伍需分享實作程式及設計文件，否則喪失領獎資格
- ✓ 測試結果每日只能提交2次
- ✓ 務必使用Machine Learning來進行辨識與分類，禁止使用任何人工標記
- ✓ 不可私下共享程式及特徵值，但可在官方討論區公開討論
- ✓ 不可使用非主辦單位提供的訓練及測試資料集

# 競賽資料說明

Overview Leaderboard Download D

資料集

Download policy\_claim.zip

訓練集

Download training-set.csv

測試集

Download testing-set.csv

- 從一月到十二月的保單共 351273 個保單代號
- 檔案說明:
  - 保單 (policy) 資料
  - 理賠 (claim) 資料
  - 訓練集
  - 測試集, 用來上傳預測結果
- 檔案格式: CSV
- 詳細欄位說明請參照 [VariableExplanationV2.xlsx](#)
- 以 Policy Number 來串聯不同表單
- 預測下一年度簽單保費 (Next Premium)
- 檔案編碼: utf-8

# 競賽資料說明

---

➤ 來看一下資料

# 心得分享

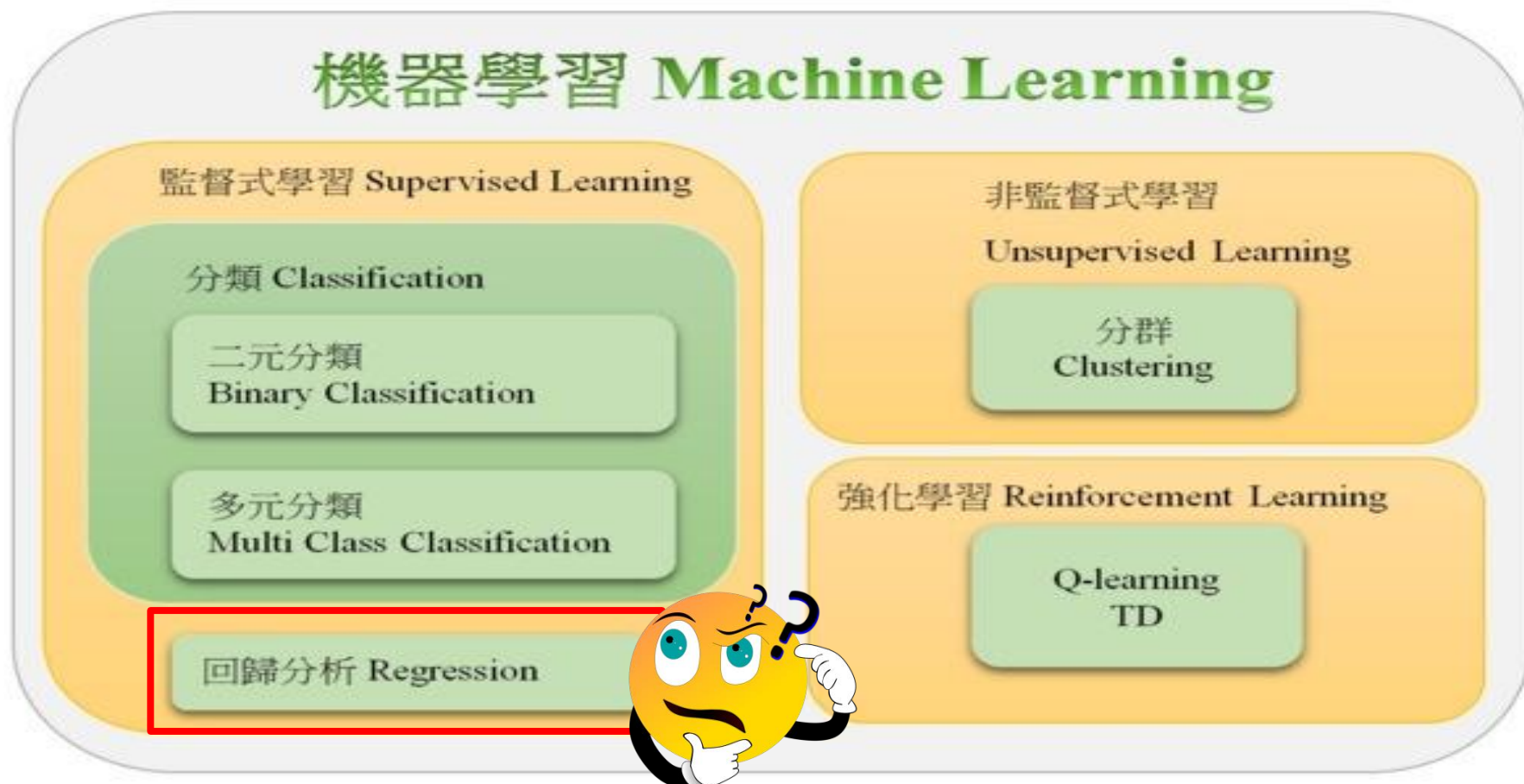
---

- 很有臨場感
- python的功力會更上一層樓
- 更了解整個機器學習的流程



# 心得分享

## ➤ 定義問題



# 心得分享

## ➤ 資料處理/特徵萃取

- ✓ 筆數不一致? 資料整併
- ✓ 處理na的值
- ✓ 性別, 婚姻....處理及轉換
- ✓ 生日 -> 年紀 (mean, 9-104?)
- ✓ 計算車子的年齡
- ✓ 廠牌車型代號 LabelEncoder
- ✓ 計算該年度的保險金額
- ✓ 計算各保險種類的保額
- ✓ 理賠次數, 金額, 自負額

# 心得分享

---

## ➤ 機器學習

- ✓ 分類

- ✓ 回歸

- ✓ `callbacks.EarlyStopping(monitor='val_loss',  
patience=10,mode='min')`

# 心得分享



向他人學習

## ➤ XGboost vs LightGBM

✓ <http://lightgbm.apachecn.org/#/>

✓ <https://blog.csdn.net/luanpeng825485697/article/details/80236759>

# summary

- 資料預處理與特徵萃取,佔70-80%
- 特徵萃取,垃圾進垃圾出
  - ✓ 影響訓練結果
  - ✓ 專業領域知識
  - ✓ 創意
- 可多了解不同的演算法,各參數的意義
- 當用很多演算法訓練效果都不佳時,應回頭思考資料面的問題



