NTU-MSLab

國泰大數據競賽

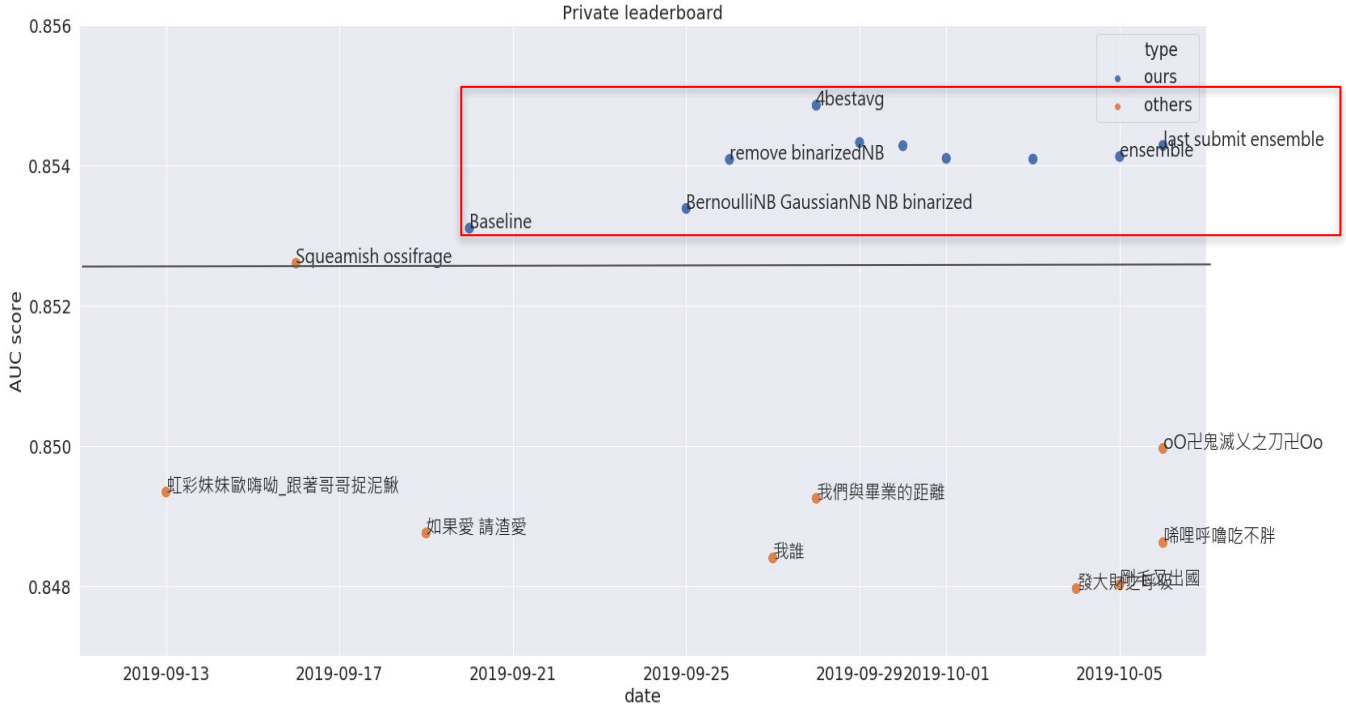
組名：import xgboost as xgb

組長：楊濟宇
組員：黃奎鈞，胡安鳳

Agenda

- 模型介紹
  - 模型表現
  - EDA & 特徵工程
  - 模型選擇
- 實務應用
  - 機器學習的角色
  - 保險業

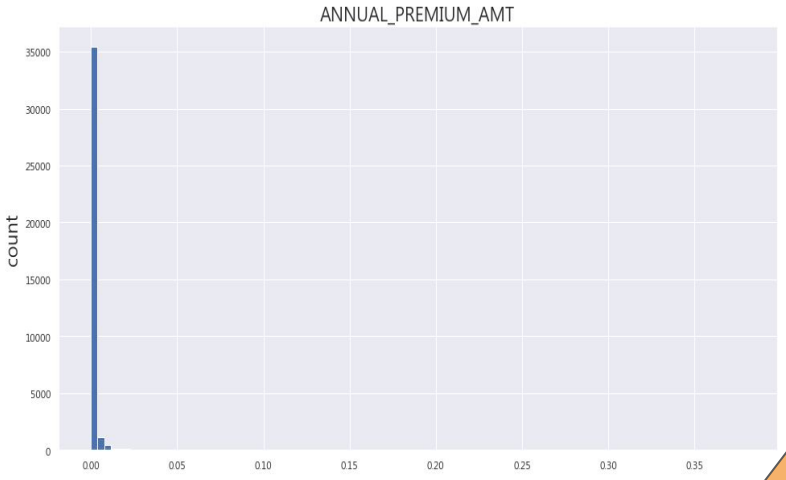
模型表現：Rank No.1，穩定高準確度

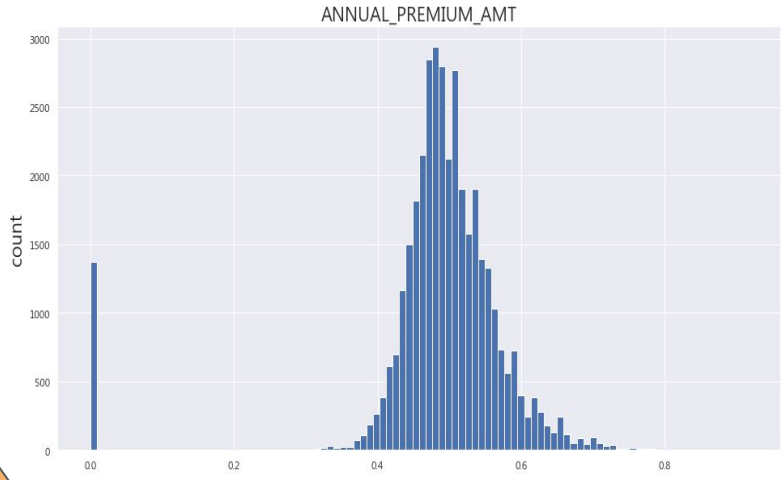


模型介紹 - EDA & 特徵工程

- Missing values (group by type)
- AMT Features
  - 22/130 ≈ 17%
  - 偏態/神秘轉換
- Binary Features
  - 80/130 ≈ 62%
- Failed Attempts

AMT Features - 偏態分析

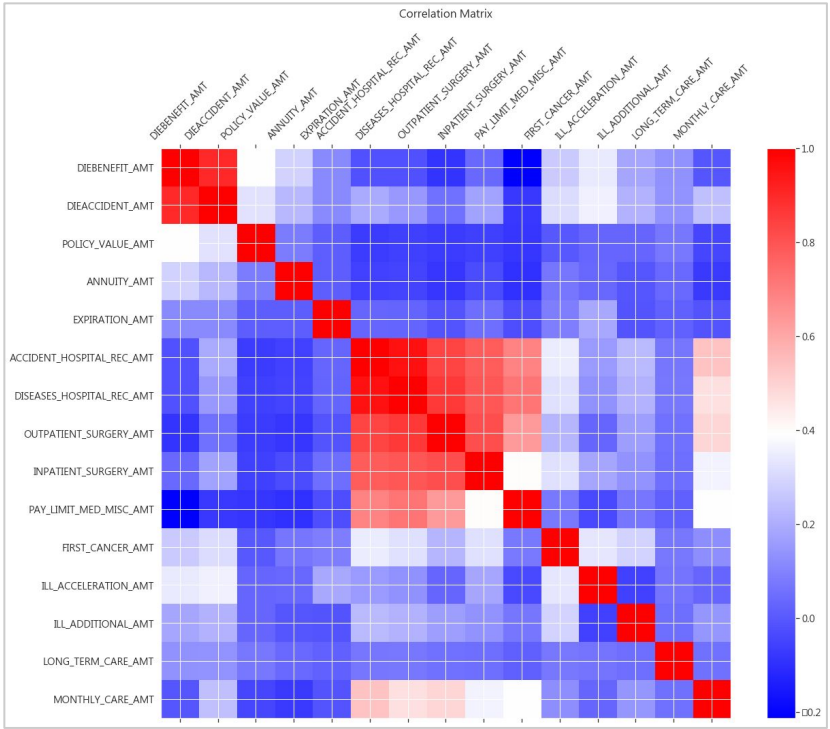




BMI 及 BANK_NUMBER_CNT 開平方

其餘開 11 次方

AMT Features - 相關性分析



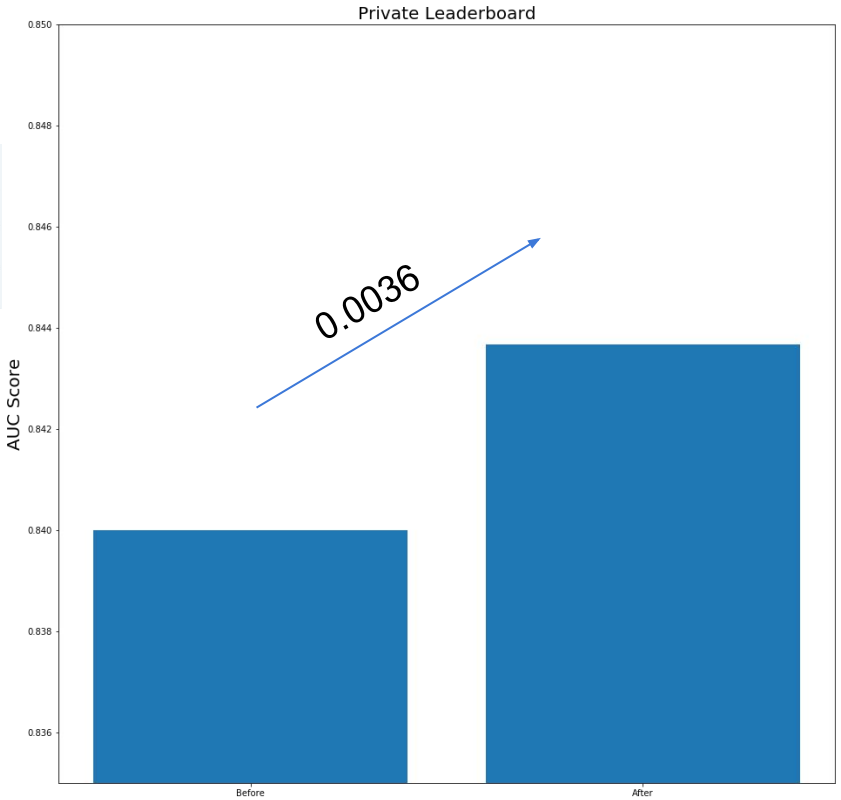
- 主成分分析 (Principal Component Analysis, PCA)
  - 2 維
  - 解釋變異
    - Dim 1 : 0.78786
    - Dim 2 : 0.11781
    - Total : 0.90567

AMT Features - 相關性分析

Public Leaderboard | Private Leaderboard

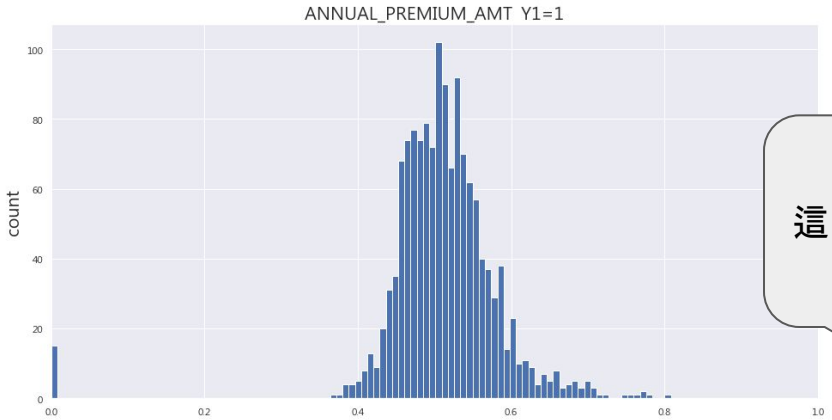
#	隊伍名稱	成員	提交次數	分數	上傳時間
1	import xgboost as xgb	3	49	0.854856	9/28/2019 11:11:53 PM
2	Squeamish ossifrage	3	26	0.852604	9/16/2019 4:34:14 PM

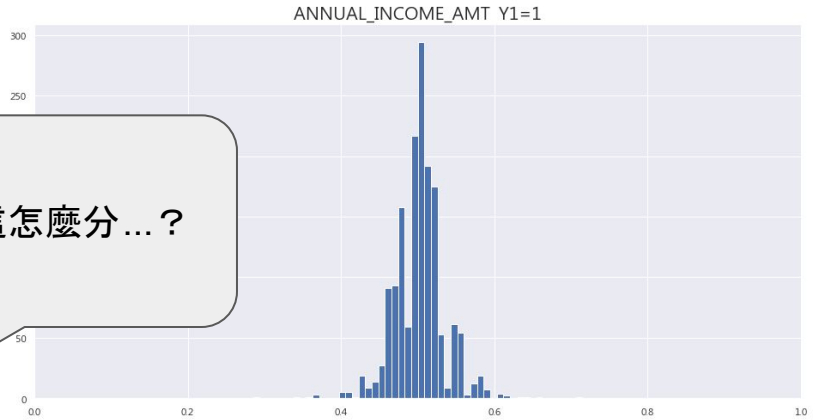
$$\Delta_{score} = 0.0022$$



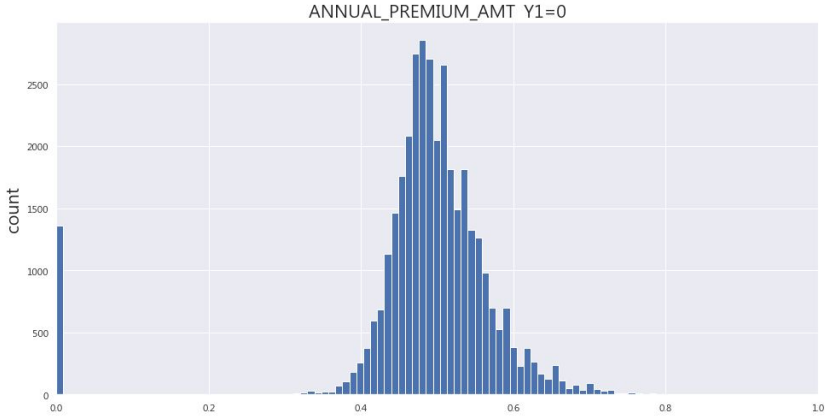
But… Can we do more?

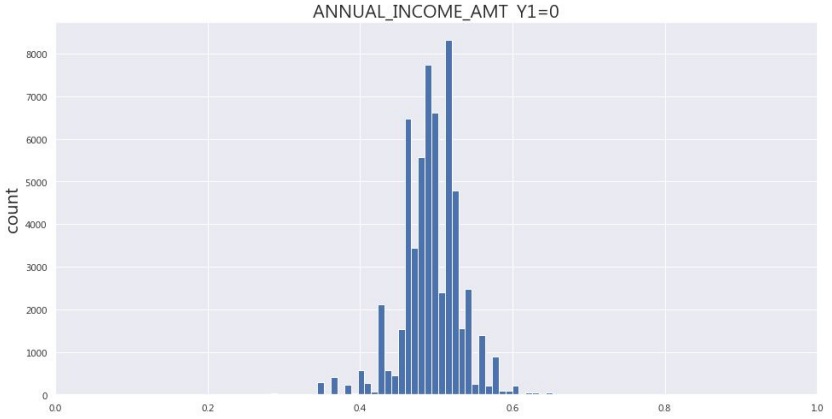
AMT Features



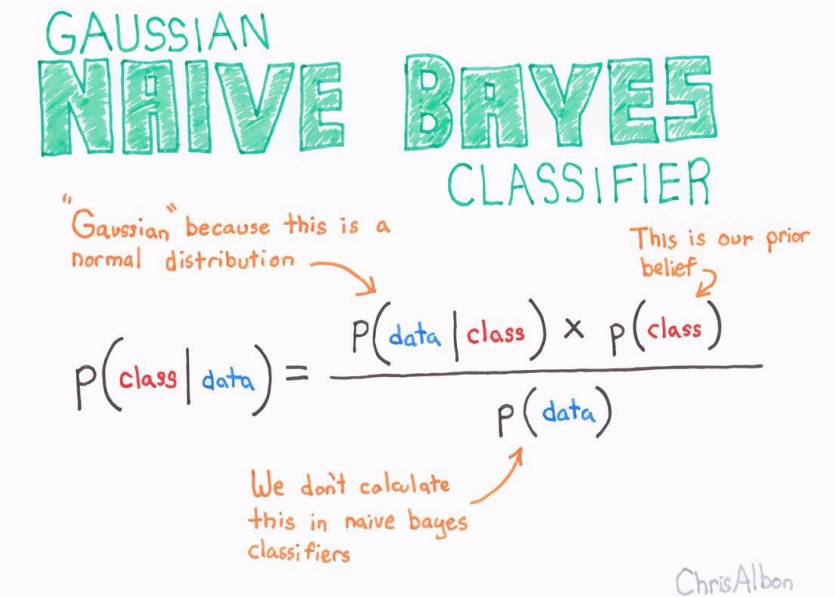


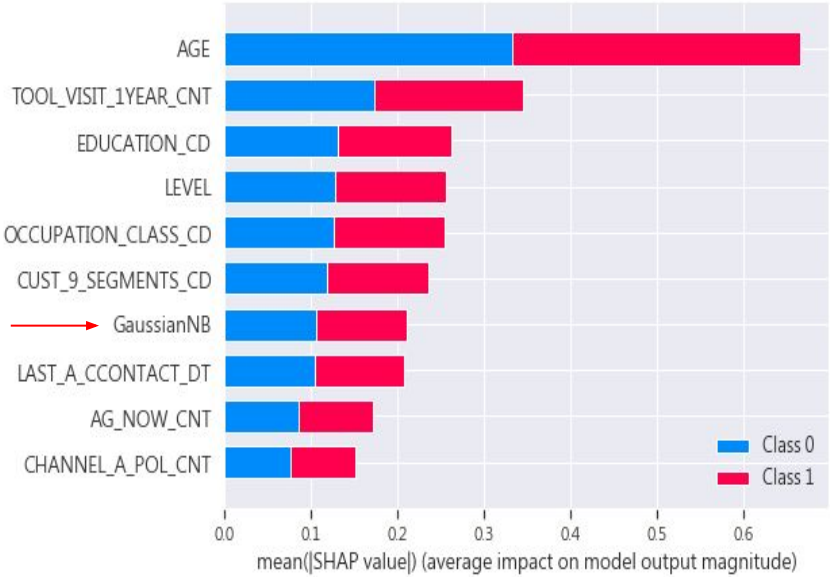
這...這怎麼分...？



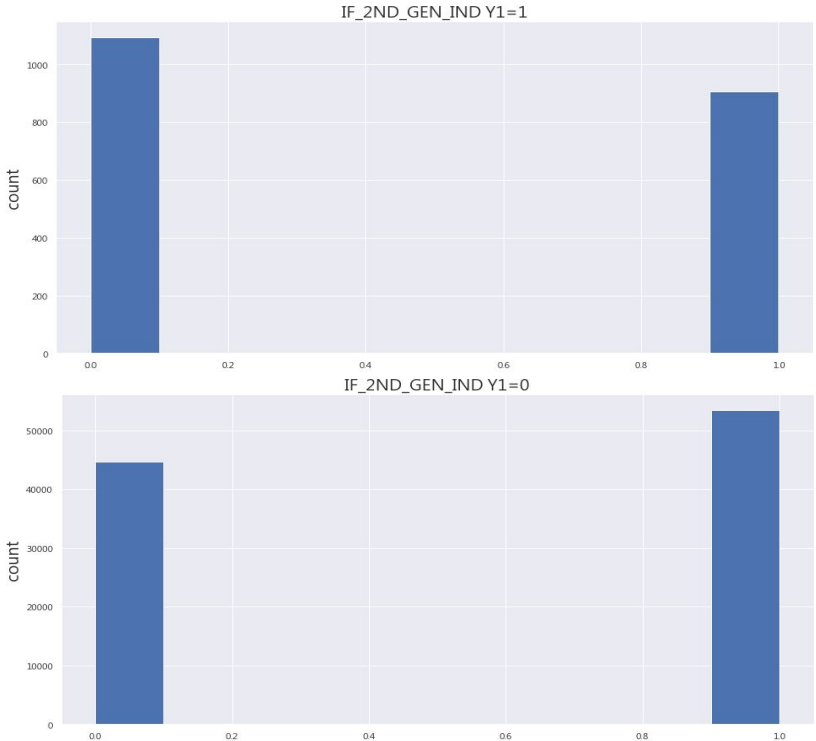


AMT Features - Gaussian Naive Bayes



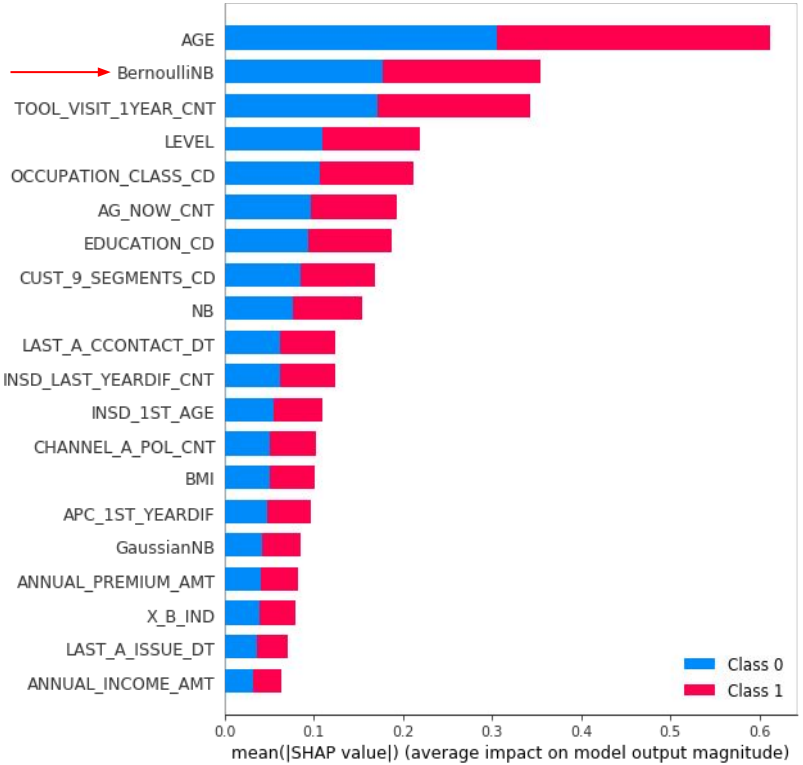


Binary Features

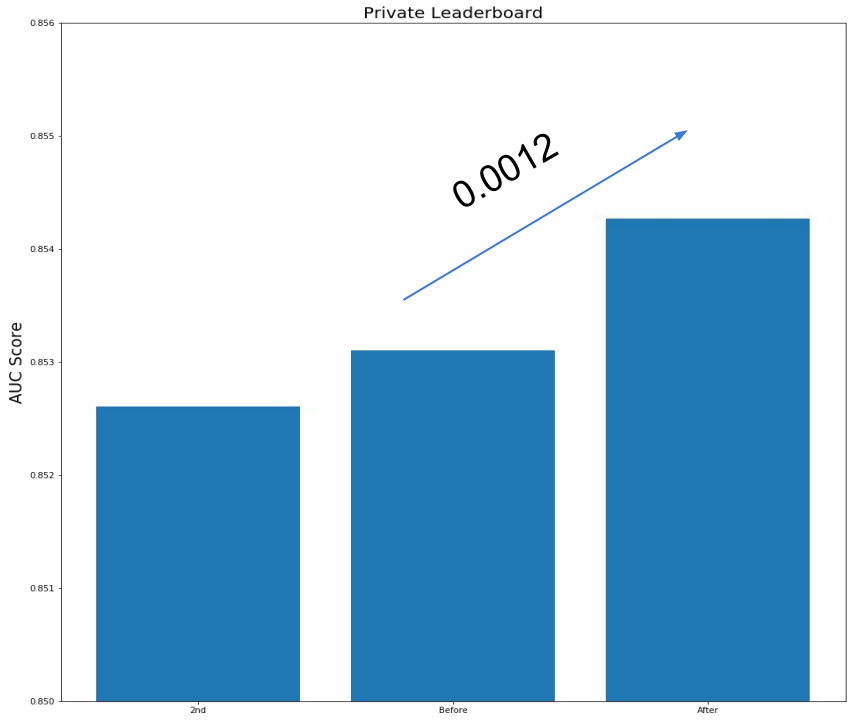


Binary Features - Bernoulli Naive Bayes

能否有效利用？



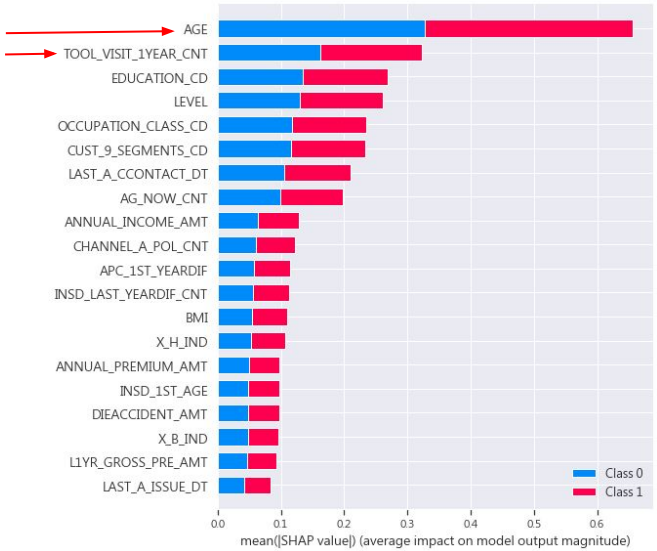
Binary Features - Bernoulli Naive Bayes

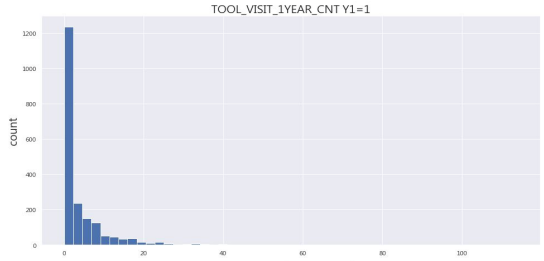


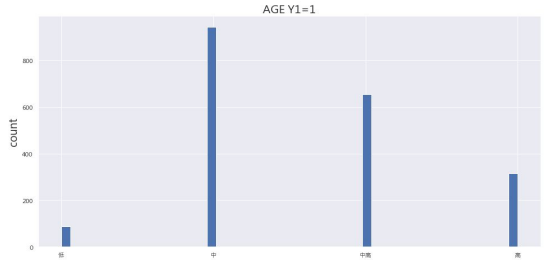
2nd

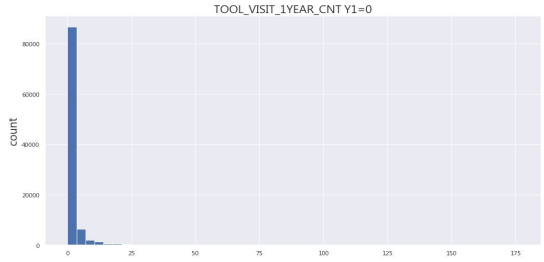
特徵工程 - Failed Attempts

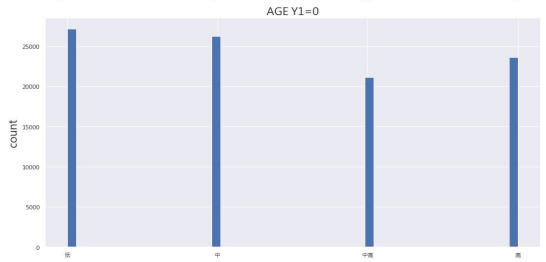
- ❑ Sum, Count, Difference, Ratio, Binarize, Binning
- ❑ Downsampling, Upsampling, Scale positive weight





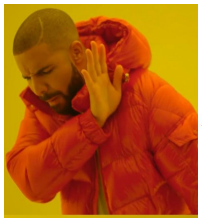






模型介紹 - 模型選擇

- LightGBM
  - 10-fold cross validation
  - 選最好的四個取平均
- 不同切法會影響訓練資料品質
  - 好的訓練資料讓你上天堂！



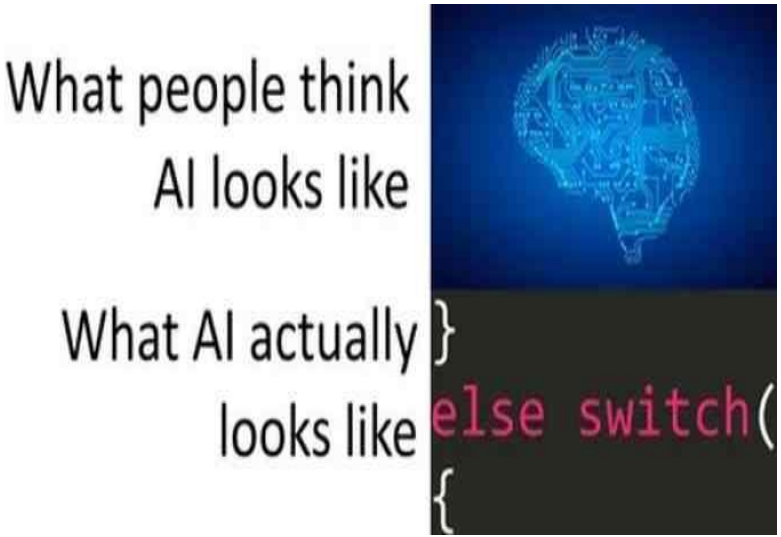
XGBoost

Microsoft
LightGBM

LightGBM	9/20/2019 11:57:28 PM	**0.855783445**	**0.8531038927**
NN	9/20/2019 6:32:22 PM	**0.8231011303**	**0.8251094491**
XGBoost	9/18/2019 8:49:20 PM	**0.8377295973**	**0.8367372172**

實務應用 - 機器學習的角色 (1/2)

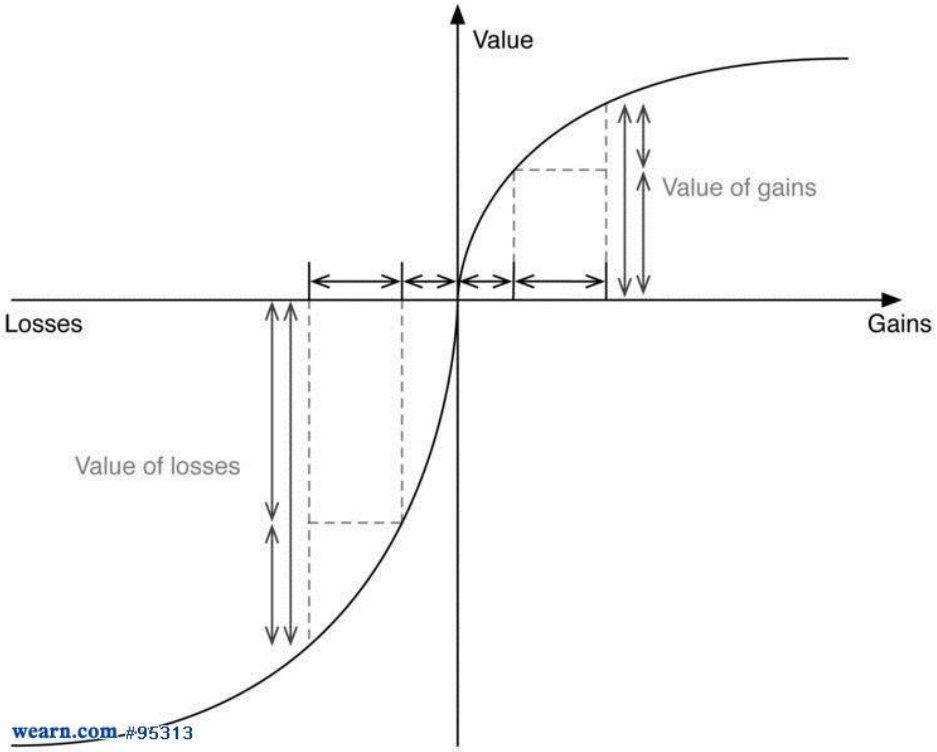
- 優缺點
  - 優點
    - 看得更全面
    - 完全理性
  - 缺點
    - 任務單一
    - 訓練資料必須乾淨



- 不是用來取代人類，而是輔助人類的！

實務應用 - 機器學習的角色 (2/2)

- 突破盲腸
- 消滅非理性效應
  - 量化風險



實務應用 - 保險業 (1/4)

- 分析業務員錯判但模型預測投保機率高的資料
- 模型與資深/績效佳業務共同決定潛在保戶
  - False Positive vs. False Negative
    - FP：成本
    - FN：獲利
- 限制：合約變異不大的保單

實務應用 - 保險業 (2/4)

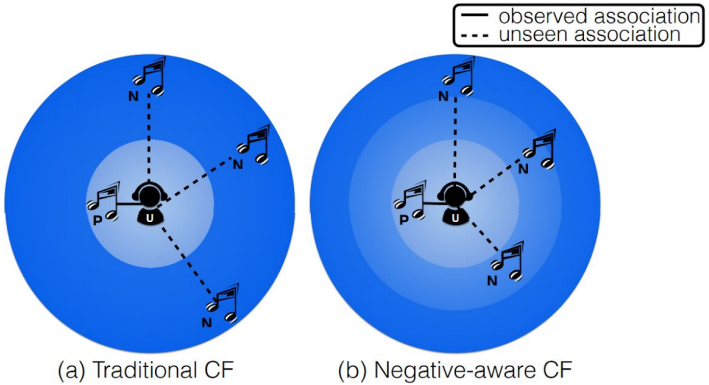
- 本題目的延伸應用之一 : 理賠預測
  - 投保人最終有無理賠
  - 二元分類問題
  - 不只降低成本也增加獲利

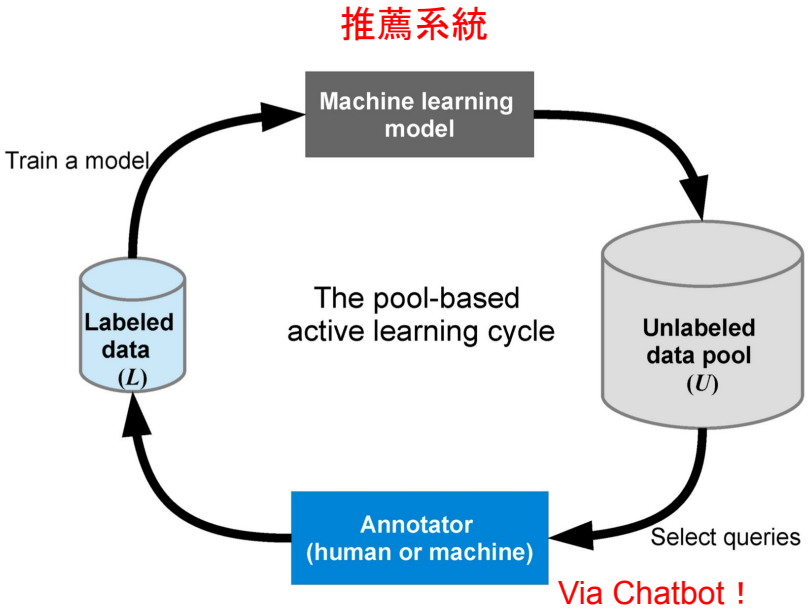
實務應用 - 保險業 (3/4)

- 本題目的延伸應用之二：推薦系統 & 聊天機器人
  - 保戶 metadata 及投保理賠紀錄
  - 推薦高機率會保且低機率理賠的保單
  - Time-aware recommendation
  - **上線後幾乎零成本還能賺更多**

實務應用 - 保險業 (4/4)

- 本題目的延伸應用之二 (續)
  : Active Learning (主動式學習)
  - 沒購買 ≠ 沒興趣
  - Chatbot





Q & A