富邦金控 Fubon Financial

客戶貸款違約風險

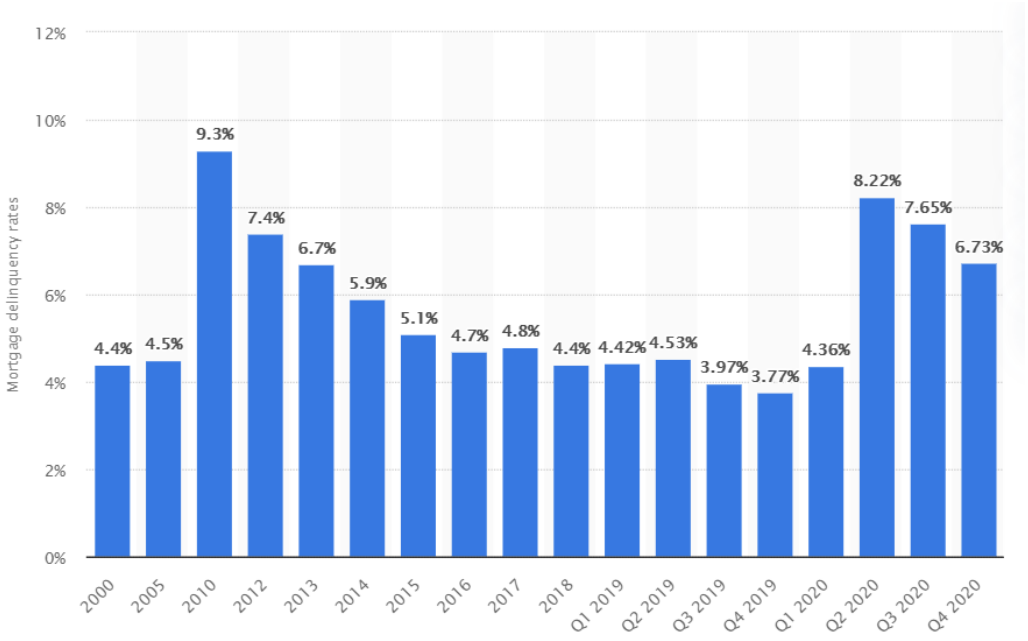
707王琥

證券-數據分析組【簡報專題】

Credit Home Default Risk Dataset

命題挑選之緣由

1. 挑戰自己
2. 數據完整、足夠、且貼近市場實況 – relational database
3. Business Model主要獲利來源
4. Wide range of applications
   - 擔保貸款 – 房屋，車貸等
   - 無擔保貸款 – 一般消費信貸
   - Margin Account



數據介紹 Introduction

Applications

- 主要數據
- 預測的變數：Target
- 當下申請貸款的基本訊息與資料
- SK_ID_CURR

Bureau

- 以前的貸款記錄、基本訊息、資料等
- 一個申請人可能有有多次的的歷史貸款記錄
- SK_ID_CURR
- SK_ID_BUREAU

Bureau_balance

- 每月餘額
- 當月的貸款狀態：進行中、已關閉、DPD: 0-30, 31-60 etc.
- SK_ID_BUREAU

Previous Application

- 與Home Credit 以前的貸款訊息。
- 一個申請人可有多此紀錄
- SK_ID_PREV
- SK_ID_CURR

POS_CASH_balance

- Home Credit 記錄的每月結餘
- SK_ID_PREV
- SK_ID_CURR

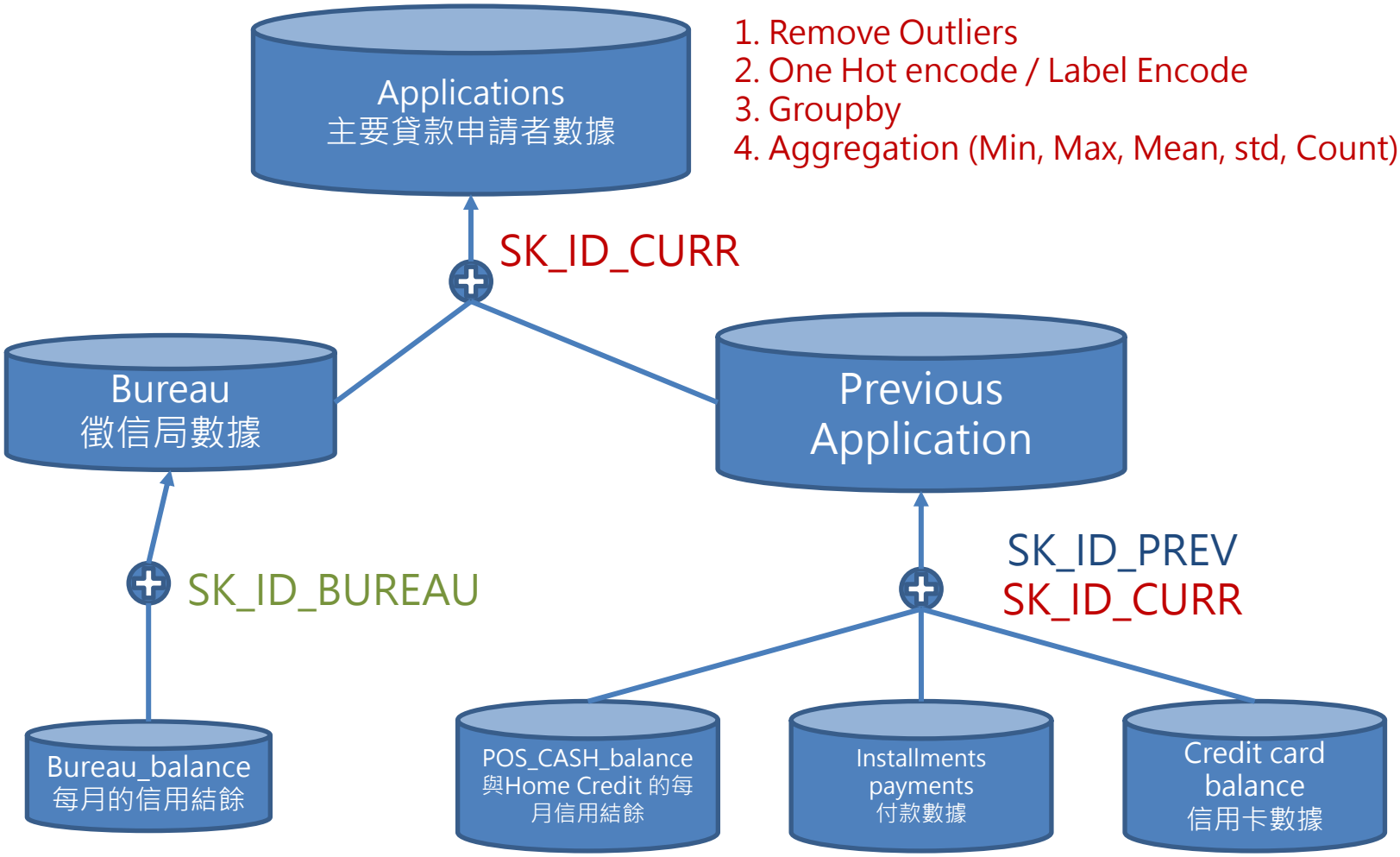
Credit card balance

- 申請人在Home Credit 的信用卡的每月消費記錄和餘額等資料
- SK_ID_PREV
- SK_ID_CURR

Installments payments

- 以前在Home Credit 的貸款付費記錄與習慣
- SK_ID_PREV
- SK_ID_CURR

Overview


Applications
主要貸款申請者數據
1. Remove Outliers
2. One Hot encode / Label Encode
3. Groupby
4. Aggregation (Min, Max, Mean, std, Count)
SK_ID_CURR
Bureau
徵信局數據
Previous Application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_CURR
Bureau_balance
每月的信用結餘
POS_CASH_balance
與Home Credit 的每月信用結餘
Installments payments
付款數據
Credit card balance
信用卡數據

Overall Steps and Idea

每個數據都要經過這個過程

1. 處理空白變數、outliers
2. 處理類別變數 – One Hot Encoder
3. 特徵工程
4. Groupby ID
5. Apply Aggregations
6. Merge with left join on ID

策略:
結合不同的研究者對特徵工程的視角，得到更好的預測和分數。

數據集中之後再進行測量與優化

7. 清理數據
8. 跑統計模型
9. 優化模型

Feature Engineering 特徵工程

Categorical

- Binary 0,1
  (性別，是否提供基本資料，是否遲繳？)
- One Hot Encode
  （工作職位？)

基本變數

- 貸款金額 – 產品金額
- 收入-貸款金額
- 收入/家庭人數
- 年收/貸款金額

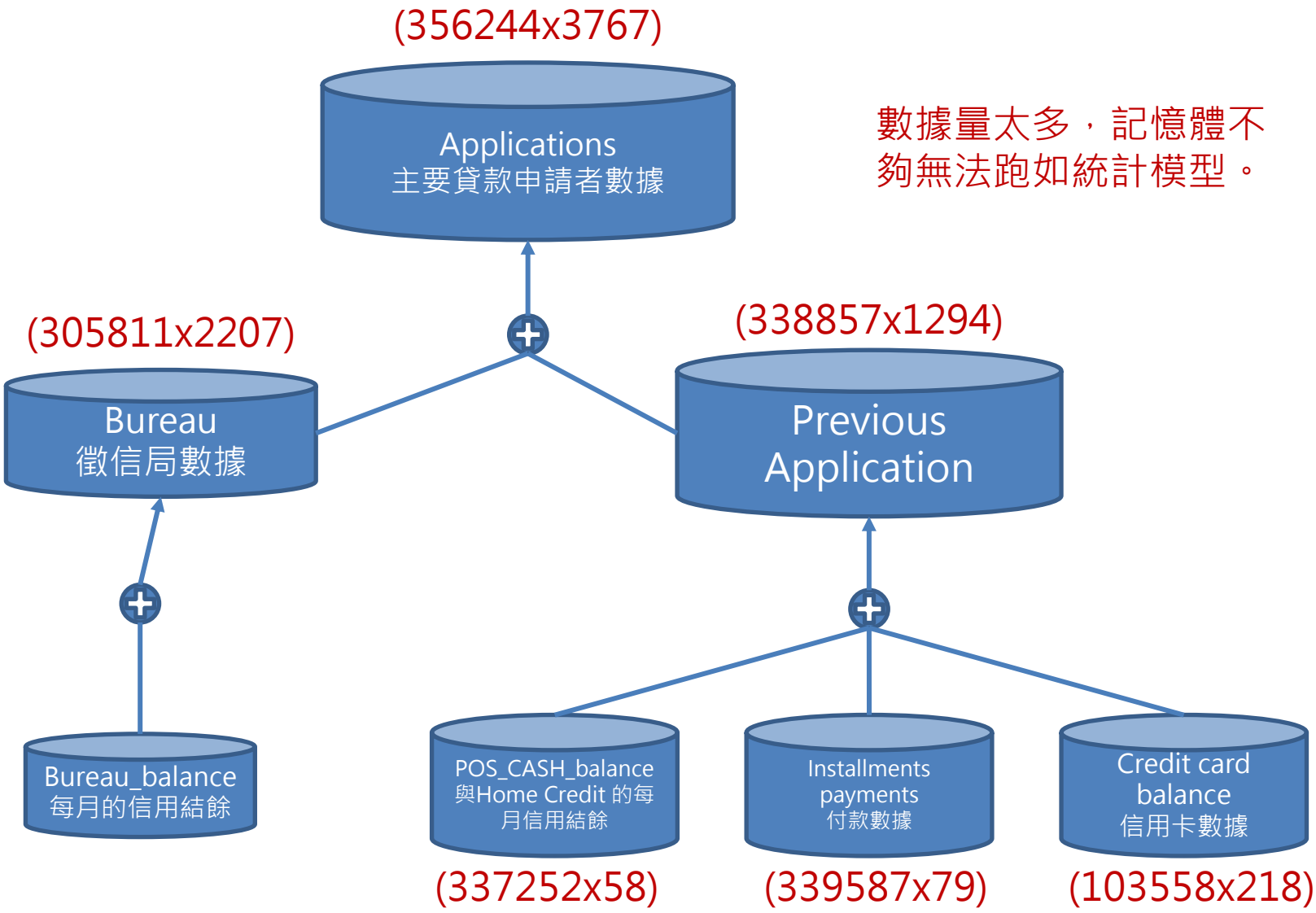
Aggregate沒有解釋的變數

- EXTERNAL_SOURCE_1, 2, 3
- EXTERNAL_SOURCE * DAYS_[]

比率變數

- 增加時間與變數之間的關係
- 貸款金額/年齡
- 貸款金額/工作時長
- 貸款剩下的比率
- 完成的貸款比率

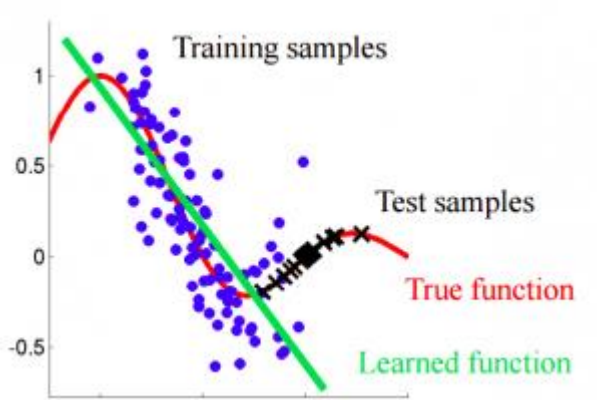
數據大小 Data Shape



Cleaning Data 清理數據

1. Reduce Memory Usage

- 找出dtype=int 的變數
- 用最大與最小值將可以改變的data type 改的越小越好
- E.g. if MAX=120, MIN=-120 change dtype to int8
- Int16, int32 改成 floating point
- 成功減少50%的memory usage



2. Delete Columns 刪除沒有訊息的變數

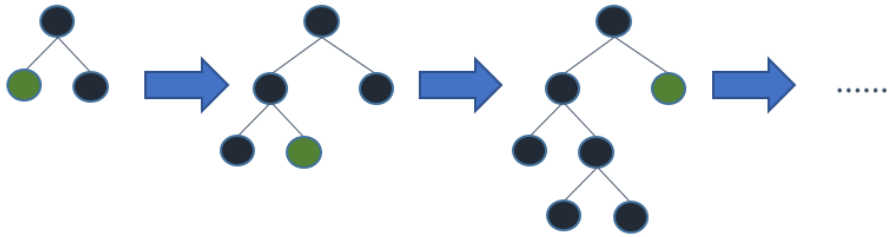
- 刪除空的變數
- 刪除0和1分佈一樣的類別特徵。(如果一個特徵有個100個0和100個1)
- 刪除Train Data 分佈與Test Data 分佈不一樣的變數 – 防止 Covariate Shift
- 刪除Feature Importance=0 的變數
- 將變數從3767 減少到 1566

LGBM Modeling

Why Light GBM?

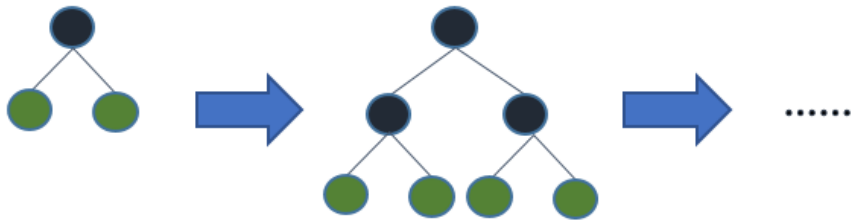
- Faster, efficient
- 不需要太多記憶體
- 可以產出相同甚至更好的準確度
- 適用於大型的數據
- 小心不能Overfit
- 限制max depth, 決策樹的層次

Light Gradient Boosting



Leaf-wise tree growth

Extreme Gradient Boosting



Level-wise tree growth

LGBM Modeling

Setup

- 5-Fold Cross Validation
- Train, Validation, Test
- roc_auc_score

Hyperparameter Tuning

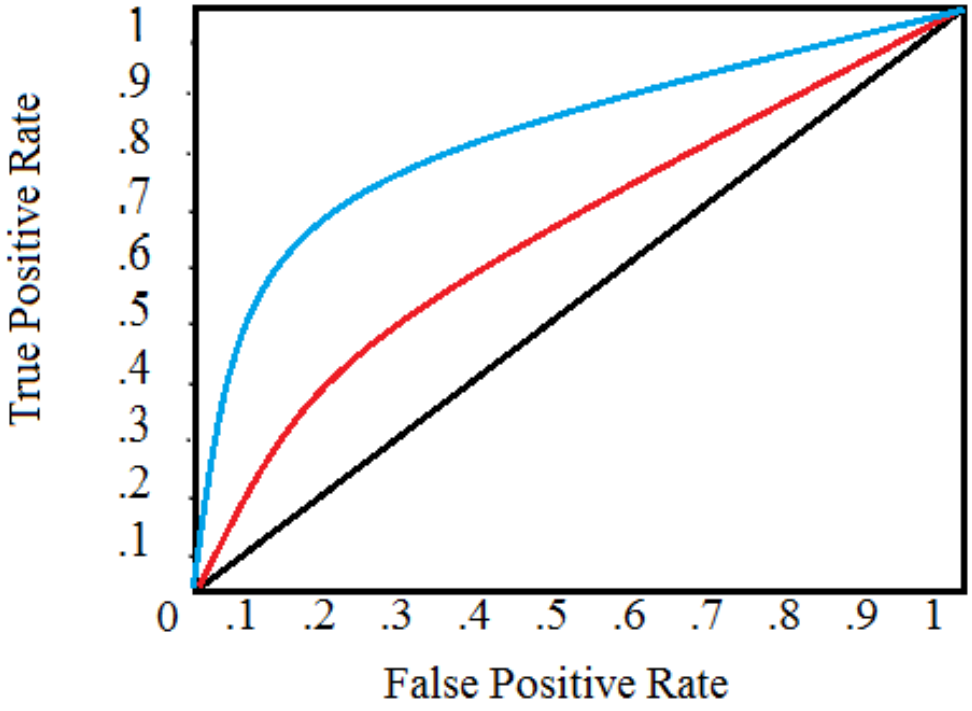
- Goss: Larger gradient
- Bayesian Optimization
- Max_depth 自己設置
- 使用 regularization terms 來防止 Overfitting (reg_alpha, lambda)
- 透過colsample, subsample 減少每一次測試的數據大小來加速模型
- min_split_gain 來加速決策樹分支
- min_child_weigh防止決策樹太多層

```
lgbm_params = {
              'boosting_type': 'goss',
              'nthread': 4,
              'n_estimators': 10000,
              'learning_rate': 0.005134,
              'num_leaves': 54,
              'colsample_bytree': .9497036,
              'subsample': .8715623,
              'max_depth': 10,
              'reg_alpha':  0.436193,
              'reg_lambda': 0.479169,
              'min_split_gain': 0.024766,
              'min_child_weight': 39.3259775,
              'verbose': -1
}
```

Results

ROC AUC Curve

- True Positive vs False Positive
- 模型預測1是1，0是0的準確度
- AUC: Area Under the curve
- 計算roc綫下的大小，約接近1越好



TP = 準確預測到target不會default的數量 0 is 0

TN = 準確預測到target會default的數量 1 is 1

FP = 錯誤預測到target不會default

FN = 錯誤預測到target會default

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=AUC%20%2D%20ROC%20curve%20is%20a,capable%20of%20distinguishing%20between%20classes.

Results – Feature Importance & score

Training Log

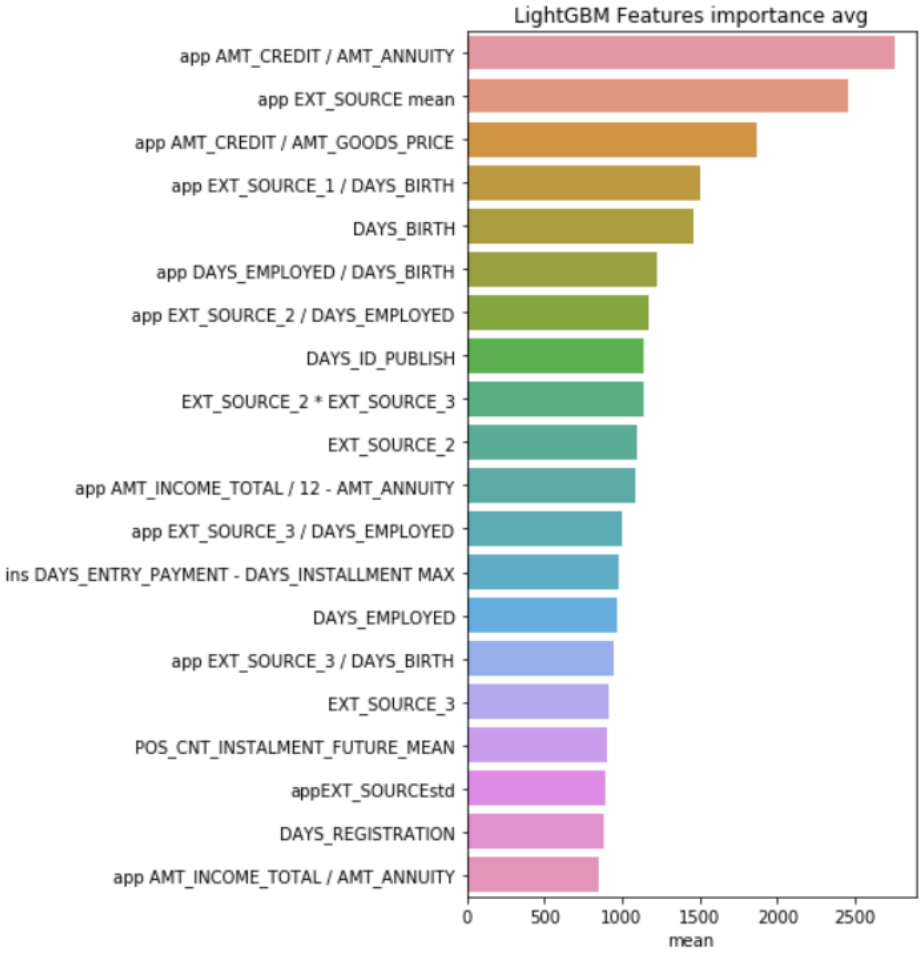
```
Fold 4 started at Mon Apr  5 10:58:06 2021
[LightGBM] [Warning] num_threads is set with n_jobs=-1, nthre
ad=4 will be ignored. Current value: num_threads=-1
Training until validation scores don't improve for 100 rounds
Early stopping, best iteration is:
[3895]  training's auc: 0.911725        training's binary_log
loss: 0.18425   valid_1's auc: 0.792797 valid_1's binary_logl
oss: 0.237642
Fold  4 AUC : 0.792797
Full AUC score 0.793690
```

Best Score

roc_auc_train	0.910625
roc_auc_test	0.793690

Submission and Description	Private Score	Public Score
prediction_0.csv 5 hours ago by Tiger Wang add submission details	0.79273	0.79680
prediction_0.csv 2 days ago by Tiger Wang add submission details	0.79367	0.79791

Feature Importance



Top 20 重要變數 – Gini Importance

1. 貸款的額度/年還款金額
2. 外部提供的訊息平均
3. 貸款額度/消費貸款中商品的價格
4. 外部訊息1/年齡（天）
5. 年齡
6. 工作年齡/年齡
7. 外部訊息2/工作年齡
8. 在幾天前更改過申請書中的個人資料
9. 外部訊息2*外部訊息3
10. 外部訊息2
11. （年收入/12 ）- 年還款金額
12. 外部訊息3/工作年齡
13. 歷史付款記錄付款日期-應該付款日期最大值
14. 工作年齡
15. 外部訊息3/ 年齡
16. 外部訊息3
17. 平均還有多少年金需要繳納
18. 外部訊息 Standard Deviation
19. 在幾天前更改過申請書的申請資料
20. 收入/年還款金額

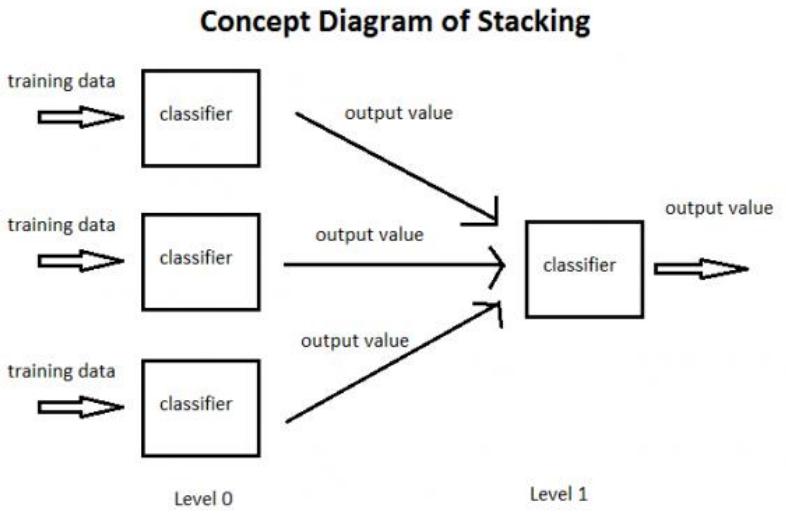
Further Research and Recommendations

1. Blender

- 使用多個不同統計模型 e.g. Xgboost, NN, Randomforest
- 記錄不同模型的predictions
- 再將各種prediction結合
- E.g. blending prediction = 0.5*predction 1 + 0.5*prediction2
- 記錄所有模型的validation set predictions
- 用這些predictions再跑一次得到最好結果的統計模型

2. 再做更多更好的特徵工程

- 瞭解更多領域知識 (Domain Knowledge)



金融場景延伸之應用

趨勢：

投資開戶的人數有再增加。

國外不介意透過借款來投資，台灣是否也會一樣？

年輕族群投資的意圖有提升，但是相對來說年輕人沒有那麼多積蓄。

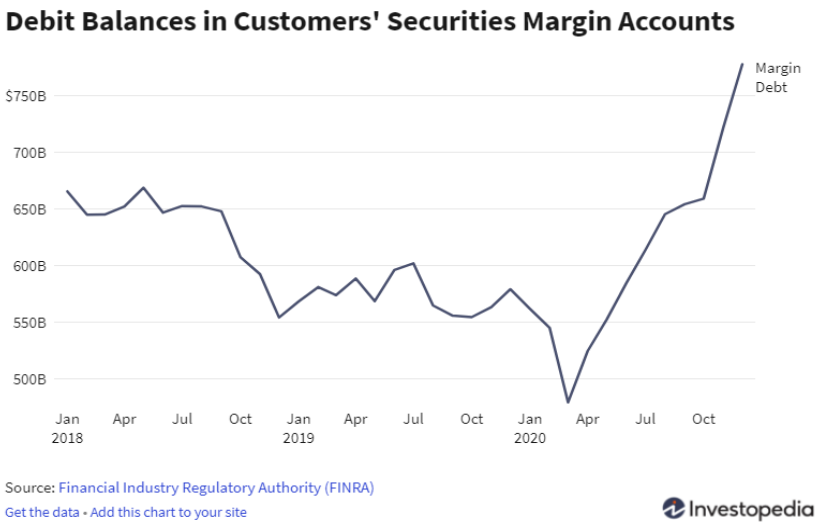


表 7、開戶人數與自然人年齡結構，暨定期定額投資金額

		2020 年	2019 年	2018 年	2017 年	2016 年	2015 年
總開戶人數(萬人)		1,124	1,057	1,024	999	977	961
新增開戶人數(萬人)		67	33	25	22	16	16
自然人總開戶人數(萬人)		1,115	1,049	1,016	992	970	954
有交易人數(萬人)		438	334	326	312	276	289
自然人開戶人數占總人口(%)		47.3	44.5	43.1	42.1	41.2	40.6
各年齡層(%)	0-19 歲	6.8	5.7	5.4	5.2	5.2	5.4
	20-30 歲	36.1	29.2	27.0	25.8	25.2	25.4
	31-40 歲	53.1	49.9	49.2	49.1	49.5	50.2
	41-50 歲	61.6	60.5	60.1	59.7	59.2	58.8
	51-60 歲	61.8	60.4	59.5	58.7	57.9	57.1
	61 歲以上	62.7	60.8	59.2	57.6	55.9	54.0
定期定額投資金額(億元)		171.3	61.5	33.8	12.1	-	-

https://www.bnext.com.tw/article/60899/broker-dealer-digital-transformation-strategy

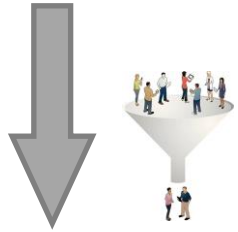
金融場景延伸之應用 – Cross Selling

- 運用各子公司客戶進行Cross selling
- 同時增加貸款利息收入、以及證券交易手續費

各子公司
客戶數據庫

匯入模型

目標客戶群分析



台北富邦銀行

- Pre-approval額度
- 開立貸款暨證券交割戶

- 富邦證券下單
- 動用計息
- 以日計息 (APR 6%)

手續費
+
貸款利率

金融場景延伸之應用 - Others

能夠預測申請人是否會Default，那是否能夠預測一家公司會不會Default呢？

- 做到更低風險的投資。

加入申請人在證券的交易數據，能否提升準確率？

- 支援臺北富邦銀行的貸款服務，讓集團整體收益得到提升。

適用於任何希望預測Binary結果的模型

- 透過此經驗能在下次遇到相同問題上有跟快速，更好的模型與做法。
- 信用卡盜刷、防詐騙、信用卡客戶轉換。

富邦金控 Fubon Financial

謝謝！
Thank You For Listening!

富邦金控 Fubon Financial

Q/A

Reference

- https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction
- https://www.kaggle.com/willkoehrsen/introduction-to-manual-feature-engineering
- https://www.kaggle.com/willkoehrsen/automated-feature-engineering-basics
- https://www.kaggle.com/willkoehrsen/intro-to-model-tuning-grid-and-random-search
- https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f
- https://www.kaggle.com/aantonova/797-lgbm-and-bayesian-optimization
- https://www.kaggle.com/jsaguiar/lightgbm-7th-place-solution
- https://www.kaggle.com/ashishpatel26/different-basic-blends-possible
- https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs
- https://github.com/KazukiOnodera/Home-Credit-Default-Risk-xgboost/#:~:text=Light%20GBM%20is%20a%20fast,many%20other%20machine%20learning%20tasks
- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=AUC%20%2D%20ROC%20curve%20is%20a,capable%20of%20distinguishing%20between%20classes
- https://www.fubon.com/financialholdings/citizenship/downloadlist/downloadlist/Fubon_CSRreport_2019_CHebook.pdf.pdf