108年亞洲矽谷產業專業人才發展基地
資料科學與大數據分析師養成班第04期

智慧型信用貸款預測系統

第一組-黃楹逸、戴碧玉、宋宏達、林子鈞、左清安、李祖賢、劉沂貞
指導老師-蔡智勇、黃登揚

專案目標
Overview

利用大數據資料
預測客戶是否可能違約，
自動給予授信決策。

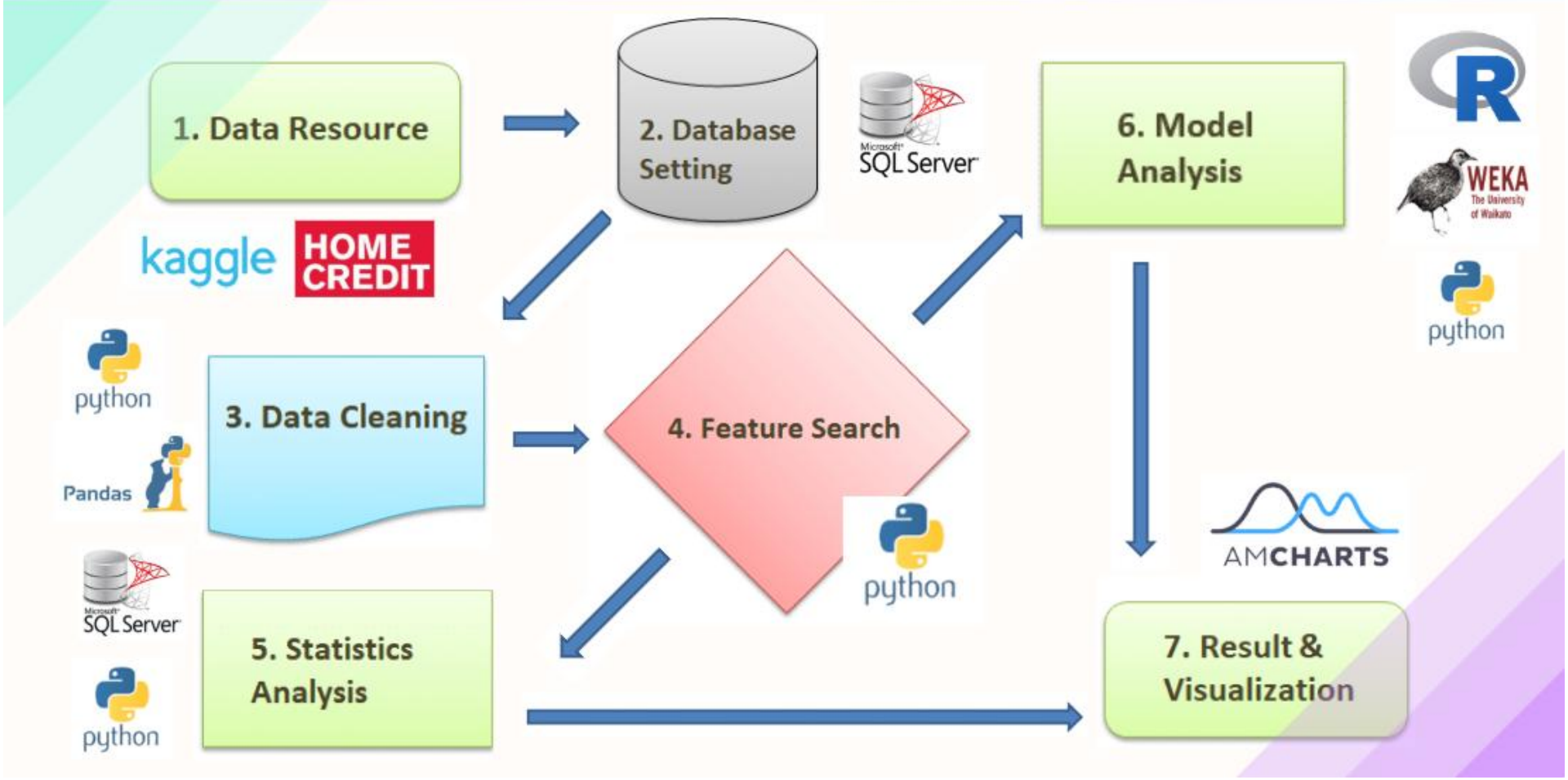
KEY PRESENTATION POINTS

在台灣，每個月有687萬貸款人、
2290萬筆貸款申請，核准率約 30%

一般信貸審核時間大約6~8個工作天，
期間需要經過照會、審核、對保
等等程序後才會正式撥款。

1. 審核費時　2. 授信決策偏誤

系統架構圖 System architecture diagram



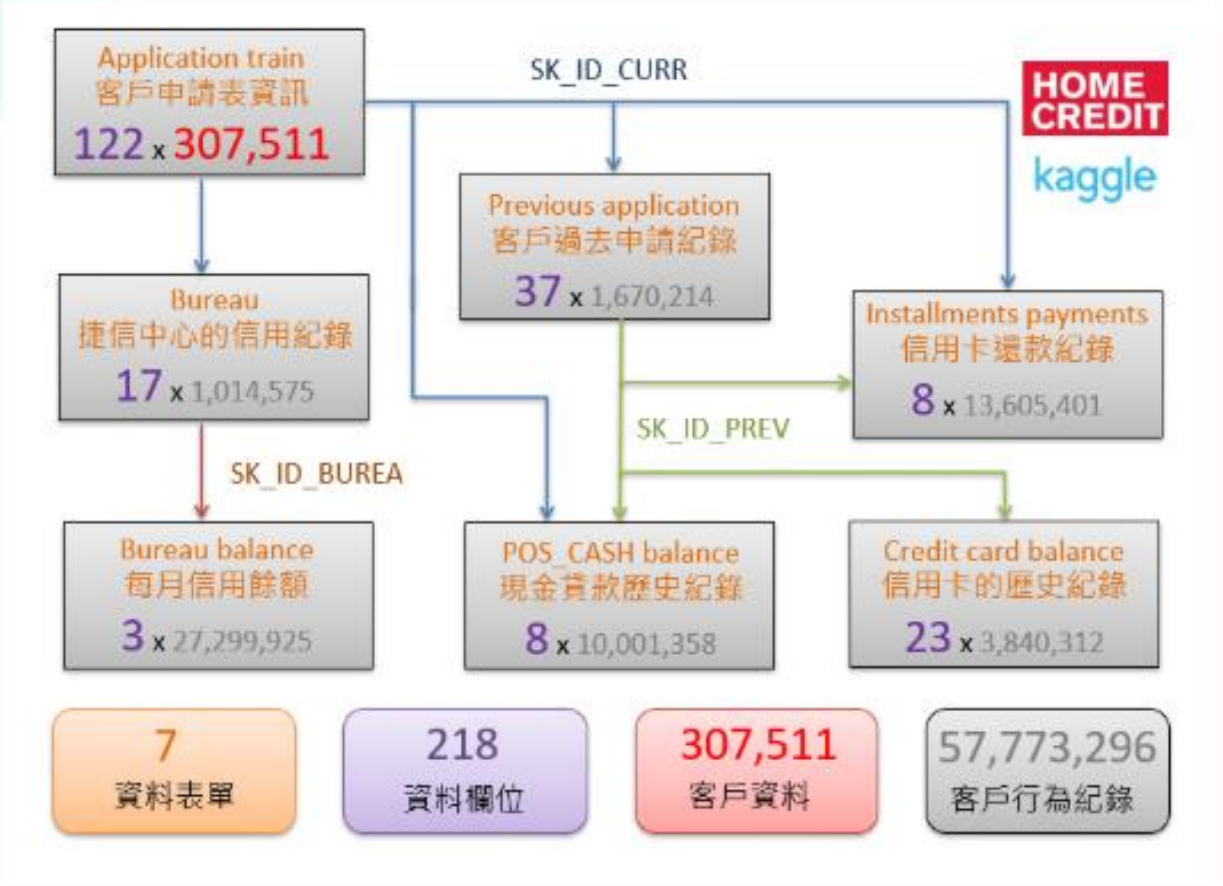
2018 DEFAULT RISK競賽

Kaggle是一個數據建模和數據分析競賽平台。企業和研究者可發布數據，由統計學者和數據挖掘專家進行競賽以產生最好的模型。

捷信集團業務

Home Credit成立於1990年，主要從事銷售貸款、消費金融、零售銀行業務等，並在俄羅斯、中國、印度等9個國家開展業務。

資料庫設定 Database Setting



7
資料表單
218
資料欄位
307511
客戶資料
57773296
客戶行為紀錄

Gradient

客戶資料簡述

與違約紀錄(TARGET)相關原始客戶資料(SK_ID_CURR)共307511位客戶，資料欄位共122個，資料內容多空值。

所以首先要梳理客戶過往的貸款行為將各表資料相互連結找，從實際業務的角度來構造每個用戶的特徵。

資料清洗 Data Cleaning

STEP 1

刪除缺失值 47% 以上的欄位
由於大量缺失先處理

STEP 2

職業欄 31% 缺失值，由於為類別型態，以NA值填滿不刪除

STEP 3

剩餘缺失值欄位為連續數值，因無法判斷，故逐列刪除
觀察離群數值資料一併刪除

資料筆數 77267
還款正常筆數 71112 備註 0
還款異常筆數 6155 備註 1

	Total	Percent
COMMONAREA_MEDI	214865	69.872297
COMMONAREA_AVG	214865	69.872297
COMMONAREA_MODE	214865	69.872297
NONLIVINGAPARTMENTS_MODE	213514	69.432963
NONLIVINGAPARTMENTS_MEDI	213514	69.432963
NONLIVINGAPARTMENTS_AVG	213514	69.432963
FONDKAPREMONT_MODE	210295	68.386172
LIVINGAPARTMENTS_MEDI	210199	68.354953
LIVINGAPARTMENTS_MODE	210199	68.354953
LIVINGAPARTMENTS_AVG	210199	68.354953
FLOORSMIN_MEDI	208642	67.848630
FLOORSMIN_MODE	208642	67.848630
FLOORSMIN_AVG	208642	67.848630
YEARS_BUILD_MEDI	204488	66.497784
YEARS_BUILD_AVG	204488	66.497784
YEARS_BUILD_MODE	204488	66.497784
OWN_CAR_AGE	202929	65.990810
LANDAREA_MODE	182590	59.376738
LANDAREA_AVG	182590	59.376738
LANDAREA_MEDI	182590	59.376738
BASEMENTAREA_MEDI	179943	58.515956
BASEMENTAREA_AVG	179943	58.515956
BASEMENTAREA_MODE	179943	58.515956
EXT_SOURCE_1	173378	56.381073
NONLIVINGAREA_MEDI	169682	55.179164
NONLIVINGAREA_AVG	169682	55.179164
NONLIVINGAREA_MODE	169682	55.179164
ELEVATORS_MODE	163891	53.295980
ELEVATORS_AVG	163891	53.295980
ELEVATORS_MEDI	163891	53.295980
WALLSMATERIAL_MODE	156341	50.840783
APARTMENTS_MODE	156061	50.749729
APARTMENTS_AVG	156061	50.749729
APARTMENTS_MEDI	156061	50.749729
ENTRANCES_MEDI	154828	50.348768
ENTRANCES_MODE	154828	50.348768
ENTRANCES_AVG	154828	50.348768
LIVINGAREA_MEDI	154350	50.193326
LIVINGAREA_MODE	154350	50.193326
LIVINGAREA_AVG	154350	50.193326
HOUSETYPE_MODE	154297	50.176091
FLOORSMAX_MODE	153020	49.760822
FLOORSMAX_MEDI	153020	49.760822
FLOORSMAX_AVG	153020	49.760822
YEARS_BEGINEXPLUATATION_MEDI	150007	48.781019
YEARS_BEGINEXPLUATATION_AVG	150007	48.781019
YEARS_BEGINEXPLUATATION_MODE	150007	48.781019
TOTALAREA_MODE	148431	48.268517
EMERGENCYSTATE_MODE	145755	47.398304
OCCUPATION_TYPE	96391	31.345545

資料清洗 Data Cleaning

STEP 4

製造新欄位: 金額用比例換算、日期用年份計算、以內外部貸款資料(bureau installments_payments)觀察最高貸款金額、最高筆數、最高逾期金額、是否為新客戶、是否為長期用戶

STEP 5

標準化欄位:連續性數值欄位，名稱前加 STD 以利計算

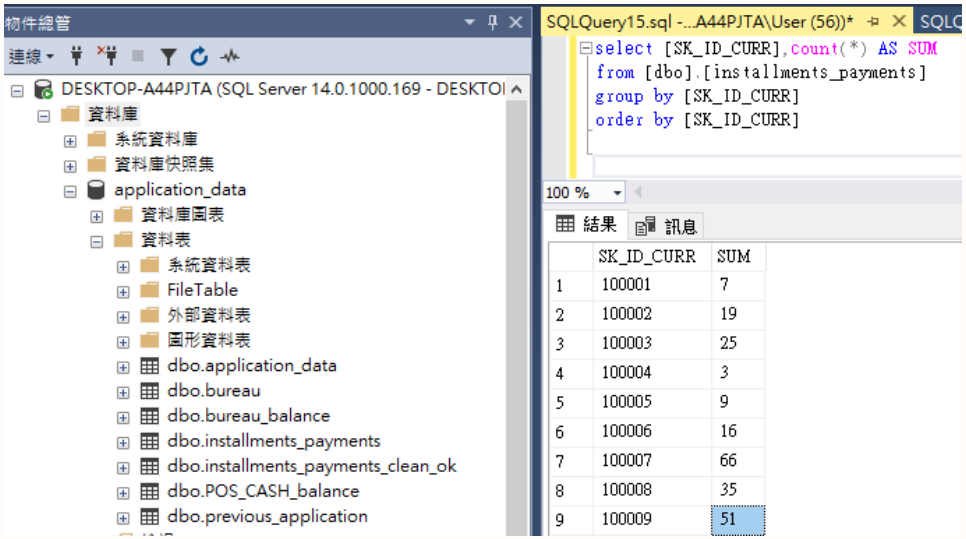
STEP 6

(11483, 83)

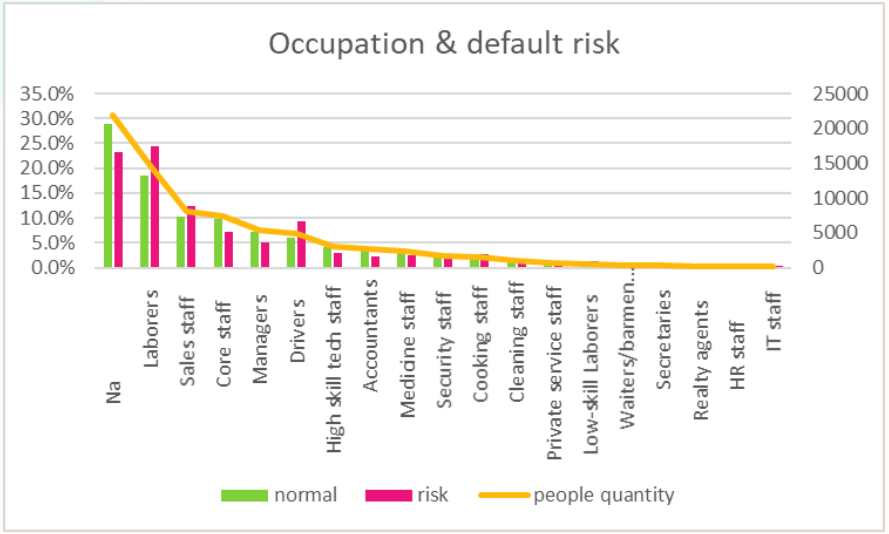
亂數抽樣:

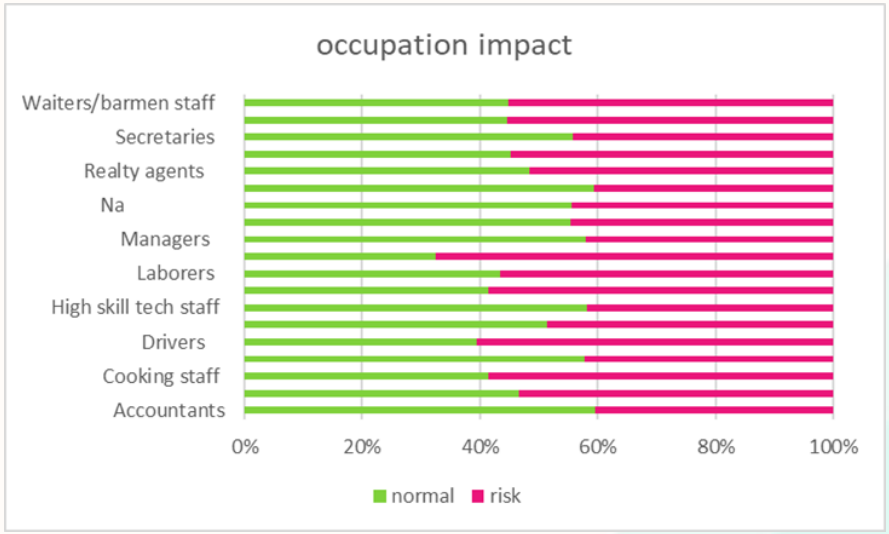
Train data 5742:5741 = 1:1

Test data 410:4520 = 1:11

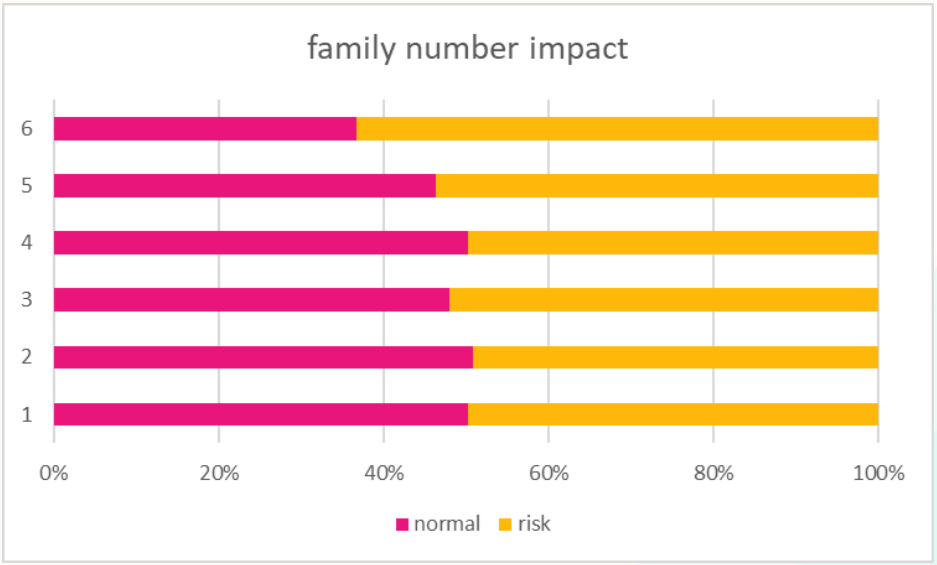


統計分析 Statistics Analysis

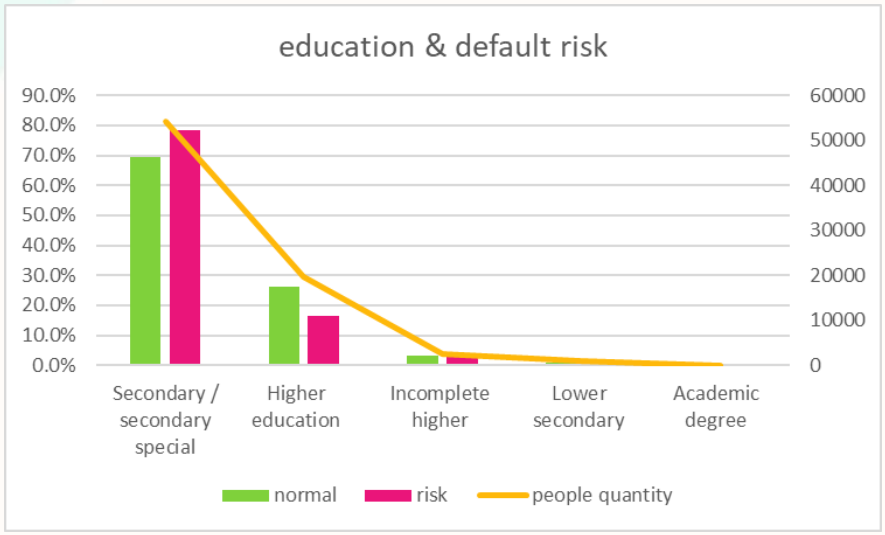


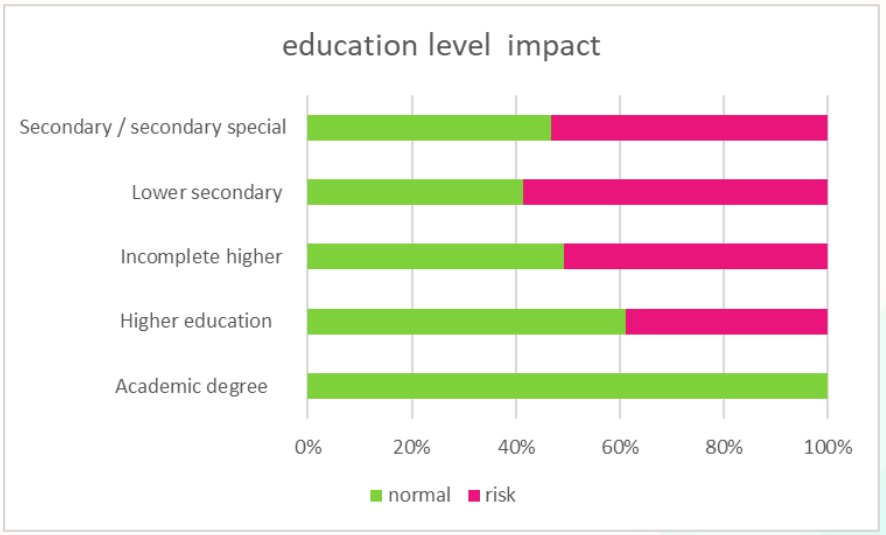


統計分析 Statistics Analysis

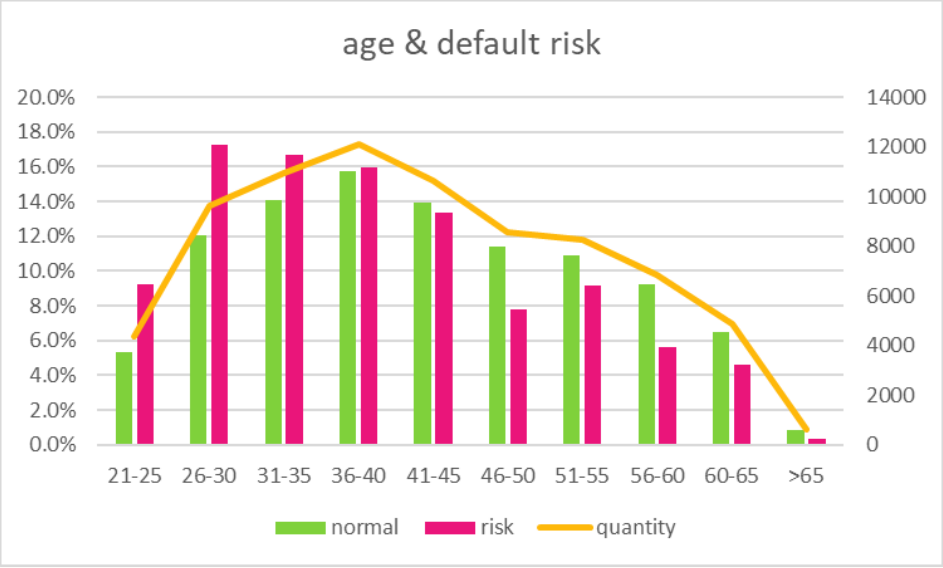


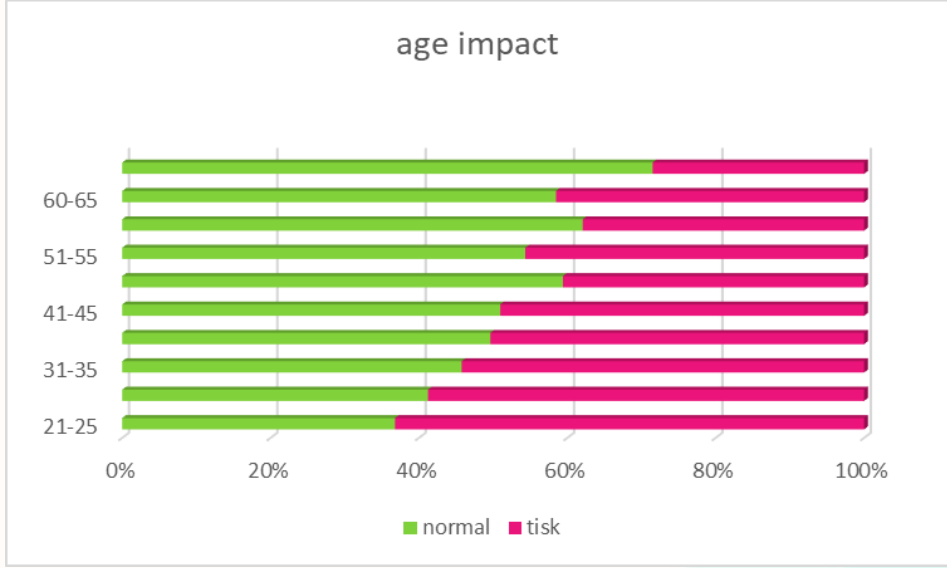
統計分析 Statistics Analysis





統計分析 Statistics Analysis





模型分析 Model Analysis

TRAIN DATA

FEATURE = 83
TOTAL DATA = 11483
TARGET 0&1 = 5:5

TEST DATA

FEATURE = 83
TOTAL DATA = 4930
TARGET 0&1 = 11:1

SimpleLogistic

R

J48

RandomForest

LightGBM

BaysNet

Python

模型分析 Model Analysis

model name	target=0 accuracy	target=1 accuracy	total accuracy
BaysNet	0.650	0.640	0.649
SimpleLogistic	0.693	0.660	0.690
J48	0.600	0.655	0.605
RandomForest	0.682	0.675	0.681
LightGBM	0.682	0.675	0.682

taget=1 accuracy : predict target=1/ actual target=1
taget=0 accuracy : predict target=0/ actual target=0
total accuracy : predict correct / total test number

```
=== Confusion Matrix ===

    a    b   <-- classified as
 3129 1391 |    a = 0
  141  257 |    b = 1
```

顧客是否違約預測準確率

HOME CREDIT USER ID 307511
DEFAULT USER 24825
DEFAULT RATE

12.38%

DEFAULT RISK
PREDICT ACCURACY

69%

Accuracy	Precision	Recall	Fmeasure
0.688	0.957	0.692	0.692

公司營運過程一定會有風險，但如何作「風險控管」就是公司最大議題！

Result & Visualization

關鍵風險因子

- A. 外部匯入資料:
  EXT_SOURCE 1.2.3
- B. 個人條件:
  居住地條件 是否提供工作電話
- C. 申請條件:
  居住地與聯絡人資料不匹配
- D. 過去貸款紀錄:
  親友間有違約紀錄　貸款比數

```
Apriori
=======

Minimum support: 0.75 (4608 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 1

Best rules found:

 1. FLAG_EMP_PHONE=1 5560 ==> TARGET=1 5560    conf:(1)
 2. DEF_60_CNT_SOCIAL_CIRCLE=0 5443 ==> TARGET=1 5443    conf:(1)
 3. REG_CITY_NOT_LIVE_CITY=0 5421 ==> TARGET=1 5421    conf:(1)
 4. NAME_CONTRACT_TYPE=Cash loans 5232 ==> TARGET=1 5232    conf:(1)
 5. DEF_30_CNT_SOCIAL_CIRCLE=0 5208 ==> TARGET=1 5208    conf:(1)
 6. DEF_30_CNT_SOCIAL_CIRCLE=0 DEF_60_CNT_SOCIAL_CIRCLE=0 5208 ==> TARGET=1 5208    conf:(
 7. FLAG_EMP_PHONE=1 DEF_60_CNT_SOCIAL_CIRCLE=0 4922 ==> TARGET=1 4922    conf:(1)
 8. FLAG_EMP_PHONE=1 REG_CITY_NOT_LIVE_CITY=0 4861 ==> TARGET=1 4861    conf:(1)
 9. NAME_EDUCATION_TYPE=Secondary / secondary special 4820 ==> TARGET=1 4820    conf:(1)
10. REG_CITY_NOT_LIVE_CITY=0 DEF_60_CNT_SOCIAL_CIRCLE=0 4800 ==> TARGET=1 4800    conf:(1)
```

管理意涵與策略

運用各種大數據資料分析技術找出「**關鍵風險因子**」，以提供予管理階層作出最佳決策方針。

可於客戶申請表單增加更多分析資料來提升預測準確率
如: 是否申請寬限期、擔保人關係、保險申請狀況

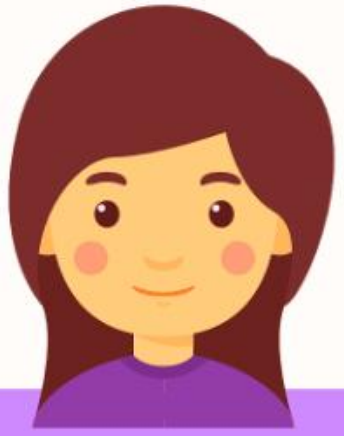
公司除了要開發與篩選出良好客戶之外，其亦建議公司可針對高風險的客戶，作風險分級的相關管理流程。

管理意涵與策略(2)

若有客戶還款困難或違約情形發生時，也能協助客戶找出問題點，共同來解決問題。

以達借貸雙方皆能雙贏的局面，再降低公司呆帳風險，並提升公司整體營業績效與淨利，且致力善盡社會責任之企業永續，促進社會經濟繁榮。

Meet the Team



黃楹逸

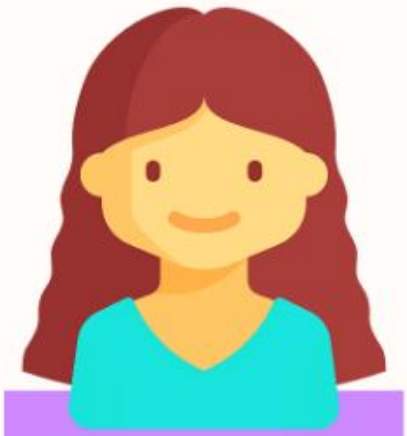
組長＆網站製作

戴碧玉

資料視覺化設計

宋宏達

資料處理與建模



劉沂貞

資料視覺化設計

李祖賢

資料視覺化設計

左清安

資料處理 &
簡報製作

林子鈞

資料處理與建模

Thank you for listening