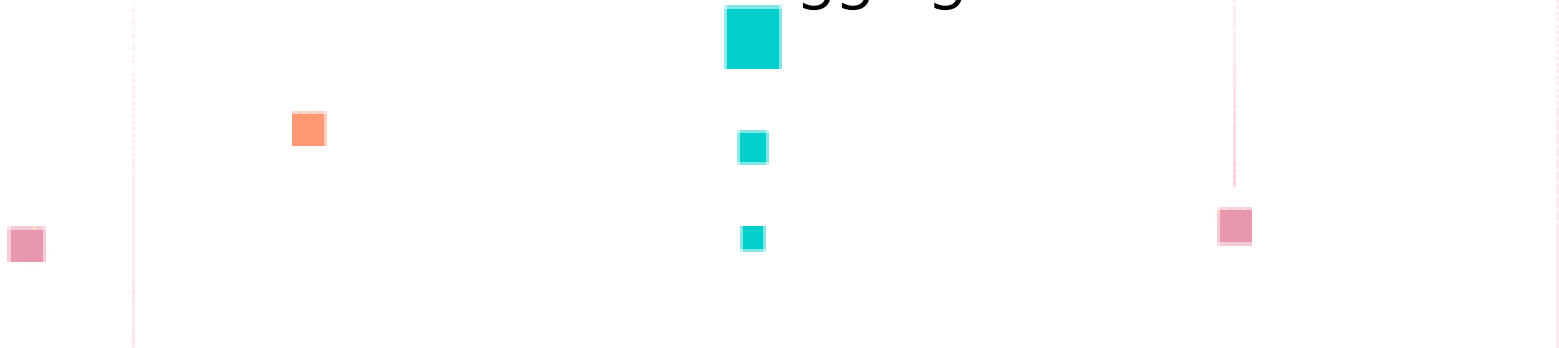
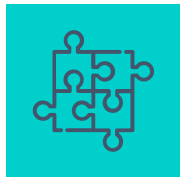
DATA SCIENCE

GIVE ME SOME CREDIT

- 任務 : 透過2年內月薪、房貸、未結清的信用貸款等特徵去分類遇到財務困境與沒有遇到困境兩種人
- 目標 : 有財務困境與沒有財務困境
- 特徵 : 房貸除以信用卡額度、年紀、月薪、債務比、未還完貸款數量、兩年內借款逾期90天、借款逾期60-89天、借款30-59天，家庭中人數(不含自己)、房貸財產抵押品數量
- 機器學習模型 : DecisionTree、Bagging、RandomForest





01

探索式資料分析

02

特徵工程

03

預測模型

1. 探索式資料分析

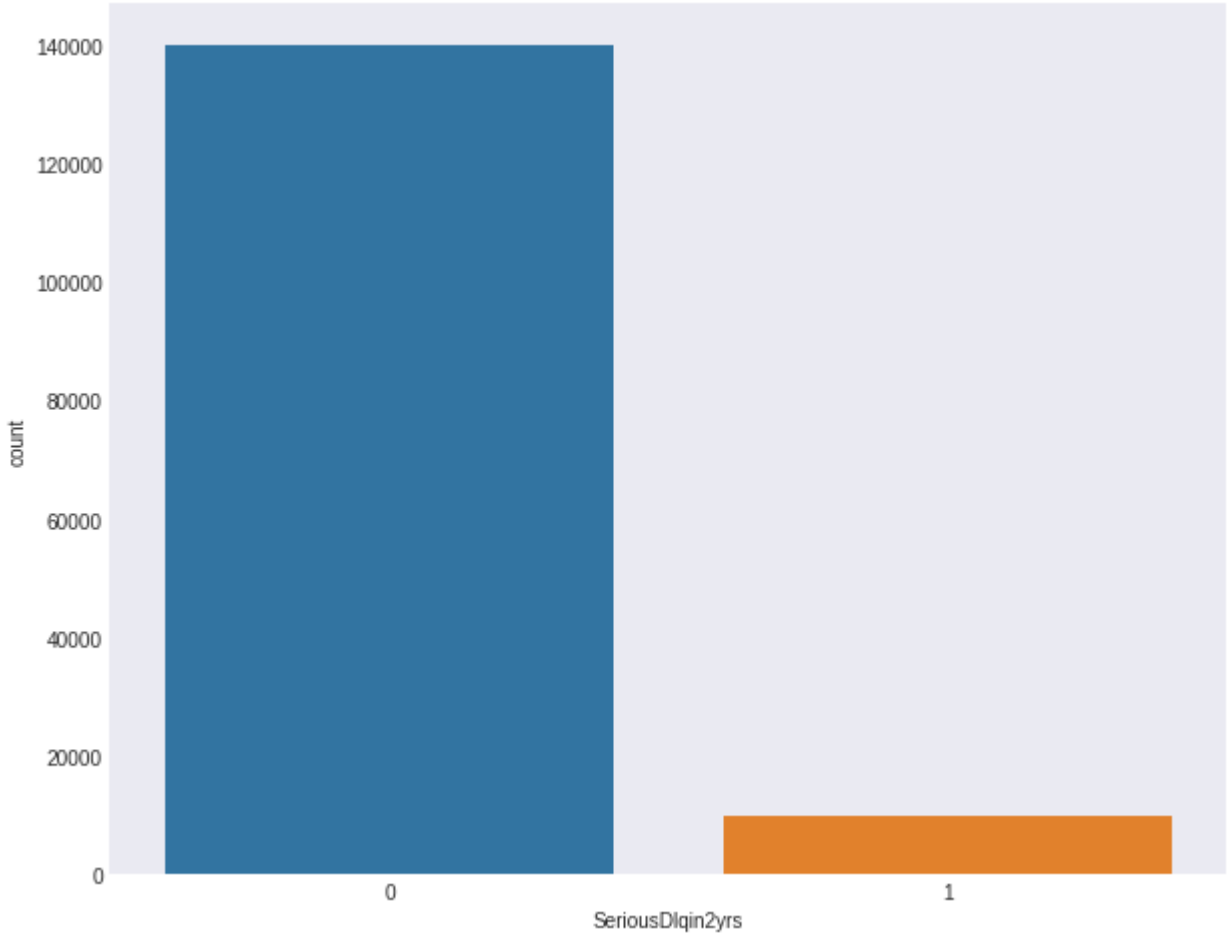
訓練資料集：

- 年齡最小值為0，為離群值，使用中位數去取代
- 家庭中人數(不含自己)特徵有很多缺失值
- 在兩年內借款逾期90天、借款逾期60-89天、借款30-59天特徵之間猜測有很高相關性
- 未還完貸款數量(ex:汽車貸款，信用卡)、房貸財產抵押品數量猜測有很高相關

測試集資料：

- 年紀最小為21 ，月薪與家庭中人數皆有缺失值

目標：有財務困境與沒有財務困境

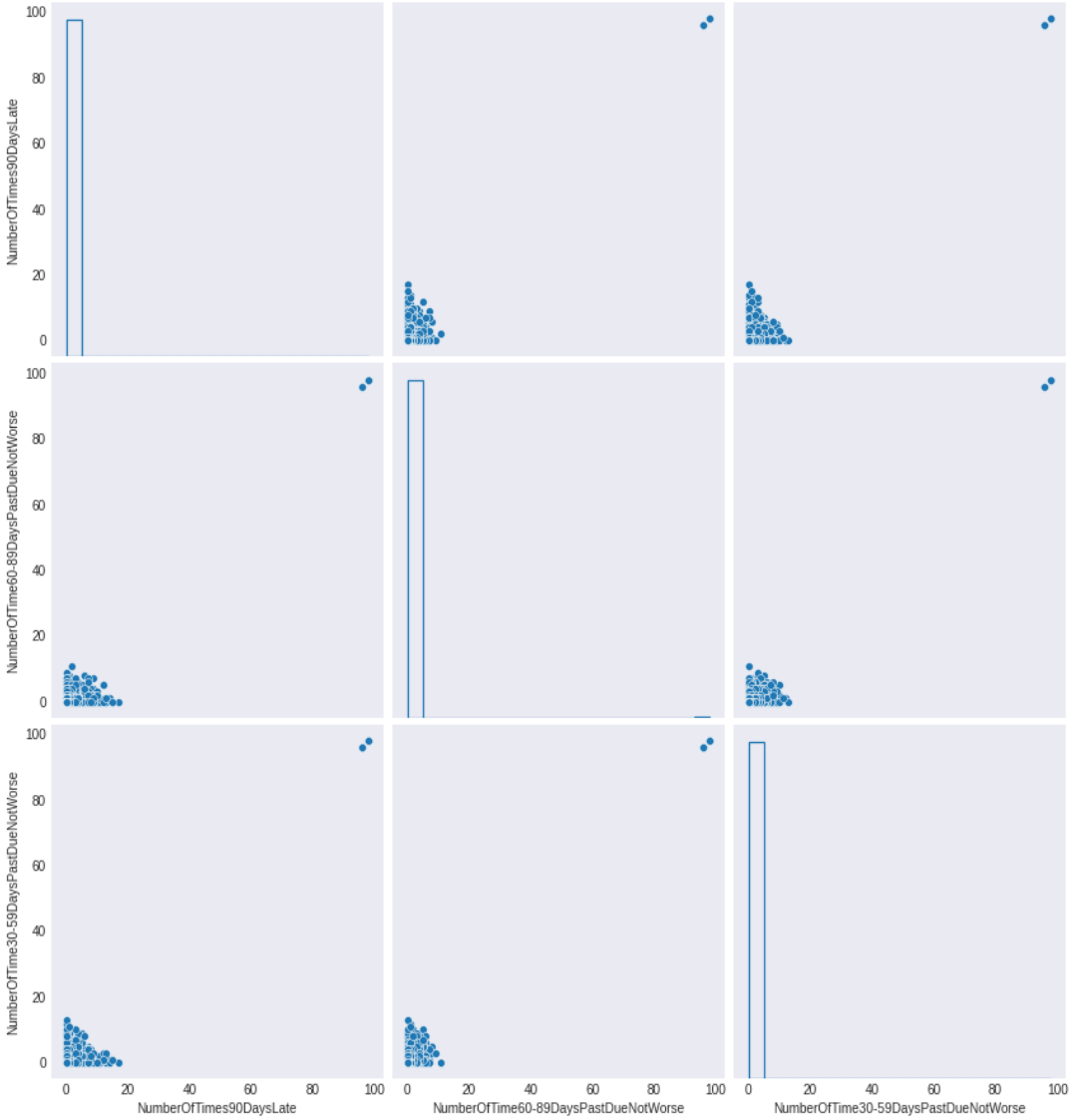


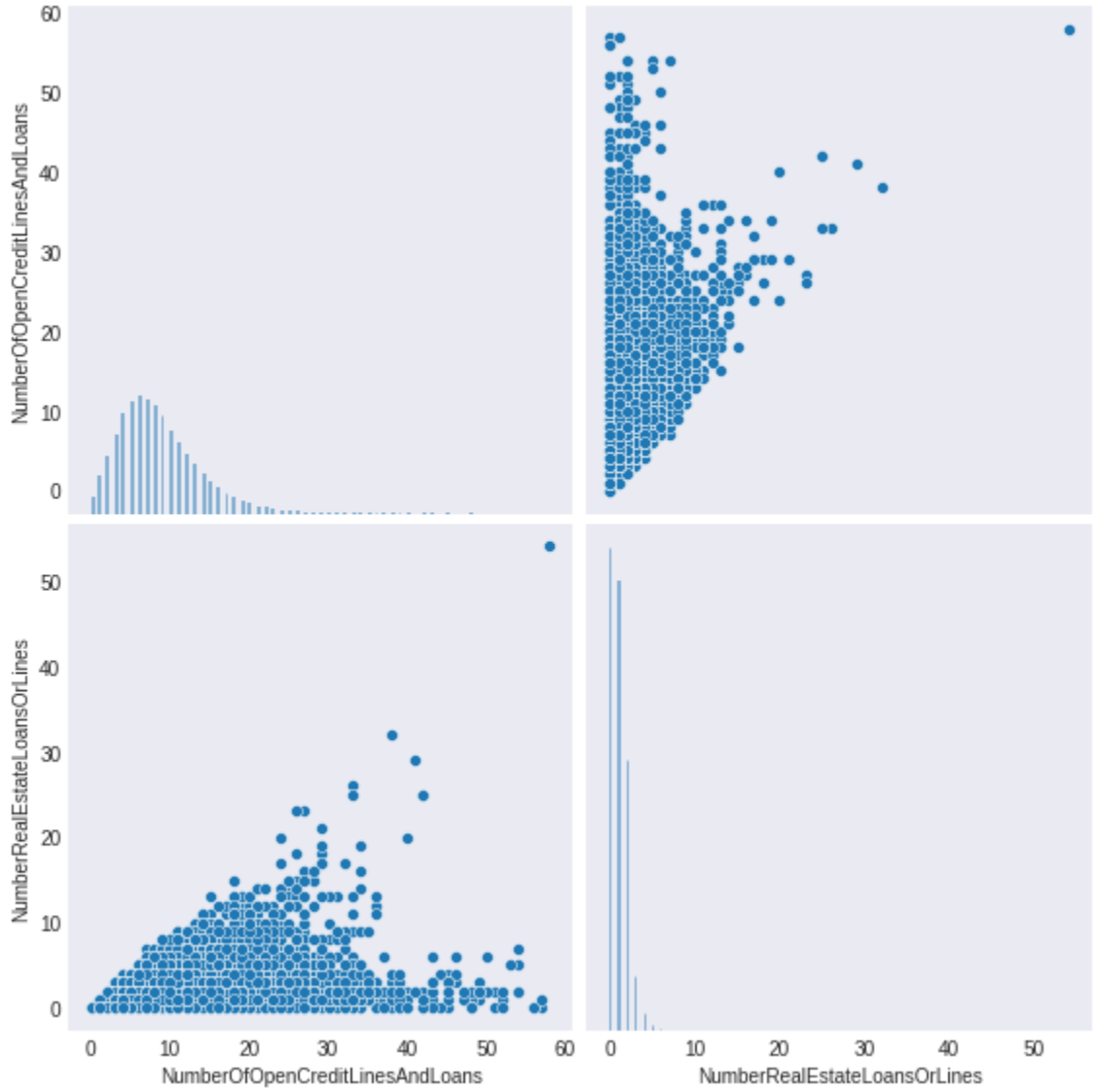
- 目標資料發現不平衡問題，可以透過集成學習模型，隨機森林可以解決
- 計算比例可得，有財務困境，有93.31%，沒有財務困境，有6.683%

2. 特徵工程

- 年紀特徵處理：將小於18歲法定年齡去用中位數取代
- 月薪缺失值用0去補植
- 將月薪透過年紀分成壯年期(18-60)、老年期(大於60)，用各自群體中位數去替代缺失值
- 在兩年內借款逾期90天、借款逾期60-89天、借款30-59天特徵合併成新的特徵，避免線性重合
- 借款逾期30-59留下作為短期或中期有違約風險重要特徵，其他兩個特徵刪除
- 未還完貸款數量、房貸財產抵押品數量合併成新的特徵

特徵成對相關圖



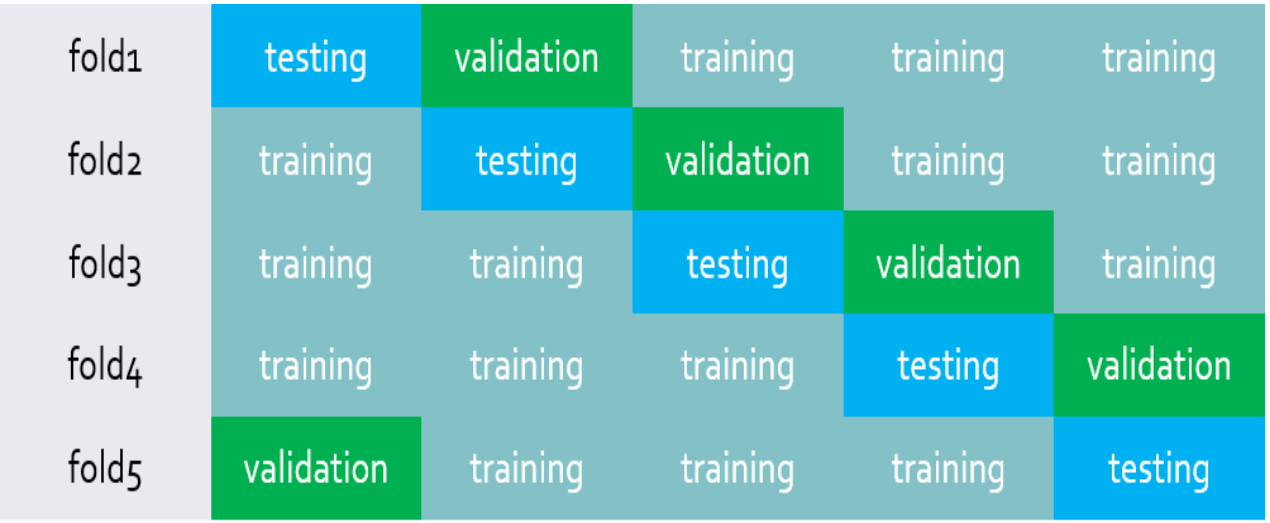


- 左圖：在兩年內借款逾期90天、借款逾期60-89天、借款30-59天特徵之間很高相關性
- 右圖：未還完貸款數量、房貸財產抵押品數量之間有很高相關

熱力圖

- 經過特徵工程後，特徵之間相關性有明顯降低

3. 預測模型



- 將訓練資料分成訓練資料跟驗證資料
- 透過5層交叉驗證，依序將資料各自為訓練資料、驗證資料、測試資料，以免發生過度擬合
- 分別使用DecisionTree、Bagging、RandomForest 機器學習模型訓練
- 將不同模型每一層驗證資料的AUC平均後比較，最高者則成為訓練模型
- 利用最佳模型去計算測試資料AUC

模型比較

Model	DecisionTree	Bagging	RandomForest
Average training AUC	0.83	0.83	0.98
Average validation AUC	0.83	0.83	0.78
Average test AUC	0.82	0.84	0.78

- 使用Bagging後，AUC相對於DecisionTree優
- 使用RandomForest後反而卻下降AUC

結論

- 理論上Bagging使用集成學習克服目標有不平衡資料，AUC可以些微上升
- 使用RandomForest卻讓AUC下降，因為採用隨機抽取特徵，容易發生過度擬合的情況
- 採用Bagging模型為此議題最佳模型