**CatBoost方法**  
**CatBoost資料預處理：**

對其他兩位元王同學整合得到的資料集，在pandas庫中調用describe查看異常資料值，並移除超過三個標準差的離群值。Train中缺失值的處理方式時直接刪去整行，test中用nan代替，使得CatBoost對其自動填補成最大值或最小值。

CatBoost是俄羅斯的搜索巨頭Yandex公司在2017年開源的機器學習庫，是Boosting族演算法的一種。和課堂教授的Ada Boost都是Tree Base的模型。它的名字來自 Category和Boosting兩個單詞， CatBoost解決了Gradient Bias以及Prediction shift的問題，從而減少過擬合的發生，進而提高演算法的準確性和泛化能力。

CatBoost嵌入了自動將類別型特徵處理為數值型特徵的創新演算法。首先對categorical features做一些統計，計算某個類別特徵出現的頻率，之後加上超參數，生成新的數值型特徵。由於原題目特徵經過我們選擇後，統一了類別型變數轉化成數字型變數的方式，所以僅使用處理後得到的資料型特徵進行分類。

CatBoost是用於多分類的Boost族演算法之一，有著獨特的處理類別型特徵使之轉化為資料類型特徵的方式。且其內嵌函式，解決了梯度偏移的問題，不容易產生過擬合現象。

NN神經網路並行處理能力強，觀察到本次訓練資料集有著大量資料特徵，類別型特徵。且其中的部分資料型特徵由於種類較少可以看做類別型，因為NN的分散式存儲和學習能力強有利於綜合利用這些特徵。另外，NN的處理結果具有一定的魯棒性。

**CatBoost參數選擇**

‘iterations’: 1000,

‘learning\_rate’:0.03,

‘l2\_leaf\_reg’:3,

‘bagging\_temperature’:1,

‘subsample’:0.66,

‘random\_strength’:1,

‘depth’:6,

‘rsm’:1,

‘one\_hot\_max\_size’:2

‘leaf\_estimation\_method’:’Gradient’,

‘fold\_len\_multiplier’:2,

‘border\_count’:128,

其他參數選擇預設值

出於對模型適用性的考慮，在F1評分的評判體系下，加入懲罰因數，使得得到的test結果更多的是no churn。這是因為銀行期望更多的人繼續他的業務。如果一個用戶希望繼續業務卻被錯分類成離開相比一個用戶不想繼續業務卻被分類成保持業務要嚴重。  
可以類比去醫院看病的情形，寧可有病，不願意有病但是醫生說沒病從而耽誤治療。

在Kaggle上最好的結果是0.22

**用NN進行分類**

**NN資料預處理**

對缺失值填補固定值進行資料預處理，接著把用隨機森林分類出的排名靠前的15個特徵和從location與population資料集提取的兩個特徵，劃分為類別型和連續型指標。接著對連續型指標，對類別型指標做標籤編碼，對連續型指標做標準化處理。

連續型變數譬如age，可以被標準化在一個區間。  
類別型變數譬如gender，做資料標籤化處理。

**NN參數選擇**

類別型變數放進嵌入層變成10維的資料，再將這些資料放入合併層。連續型變數經過稠密處理後與之前經過改造的維度相同的類別型變數連接在一起。再經過有限次稠密處理後得到最終結果。

epochs選擇30

Batch size選256

Verbose選1

其他參數選擇預設值

除了分別繳交CatBoost和NN的submission，還對對兩個結果做了投票，綜合得到一個結果。

但由於整體兩個模型的懲罰因數選的權重較不理想，得到結果不能讓人滿意。

分工  
使用CatBoost及NN對資料進行處理

參考資料

https://catboost.ai/en/docs/