Report

* Data Preprocess:
  + Data Visualize:

在進行訓練前，我們首先將資料集做視覺化，了解到label 0和1的分佈情況約為9：1，是一個很不平衡的資料集，隨後發現某些feature存在很多離群值，有一部分的數字甚至與中心分佈天差地遠。

* + Outlier Handle:

對於離群值的處理，我們嘗試使用了3倍標準差法和四分位距法去進行資料刪除。在使用3倍標準差法時，需要考慮到資料集是否符合常態分佈，而在實驗過後發現資料並不符合，所以我們取用80 %認定為離群值，但在後續套用模型時表現並不盡人意。隨後使用四分位距法，取上下超過1.5倍四分位距做為離群值，但是發現這樣會刪除將近二十萬的data，甚至取3倍四分位距也會刪除將近十五萬的data。在考量到不能刪除太多資料的情況，我們改成將離群值clip到四分位距的上下限範圍，而就在嘗試不同位距在模型上的表現後，我們認為取5倍位距是比較適合的。

* + PCA:

因為資料集本身有35個features，在使用RandomForest預先跑過一次訓練並且查看feature importance後，發現其實接近一半的features是比較沒用的，所以我們嘗試使用PCA對資料進行降維處理。PCA大致上的概念就是將高維度的資料進行線性轉換投影到一個較低維度，但是可以維持資料變異性，保留大部分的資訊，這麼一來模型就有可能比較容易學到有用的資訊而非在不重要的features上花費工夫。

* + SMOTE:

在資料不平衡的情況下，模型很容易傾向預測數量較多的那個class，所以我們嘗試對資料進行過採樣來解決這個問題。然而若是單純對少數樣本進行複製，有可能會導致模型過擬合，所以我們採用SMOTE方式進行過採樣。SMOTE引入類似k-means的概念，對於要進行採樣的樣本，會先去找出K個鄰近點，再根據公式合成出N個樣本，這樣可以根據資料分佈去做平衡，降低了過擬合的可能性，從而提高模型的泛化能力。

* + Standardization:

透過資料視覺化的圖我們觀察發現，資料分佈並不是那麼漂亮，所以我們也有嘗試使用Z分數標準化將資料進行處理，使得離群值的影響不會這麼大，同時也可以讓模型更容易進行訓練

* Model:
  + Regression

在探索模型階段，我們有嘗試過最簡單的Linear Regression來進行預測，雖然說是分類問題，但是我們可以將output threshold設定成0.5，若是output低於threshold則label為0，反之若是超過threshold則label為1。而藉由這種方法在validation data的最佳F1-score僅有0.4，隨後我們根據觀察出label分佈比率大概為9比1這一情況，調整predict結果中第data\_size \* 0.1大的數值當成threshold，來求出最終結果。

* + Boosting

Boosting系列模型是我們此次競賽中的主要模型，其原理如課上所教，會將前一棵樹犯錯的樣本賦予較高的權重，使得下一棵樹會傾向去學到該錯誤樣本的資訊以讓其判斷正確。而Boosting系列模型我們嘗試了三個：XGBoost，LightGBM和CatBoost。三者在細節實作上各不相同，但大致上可以將LightGBM看成XGBoost的輕量版，而CatBoost則是增加了許多資料處理的自動化，減少人工調參的難度，使得模型比較不容易過擬合，也正因如此，CatBoost在本次競賽中是我們得分最高的模型。

在調參上，三者都有一個很重要的參數scale\_pos\_weight。該參數是用於處理不平衡資料集，其原理就是將多數和少數資料賦予不同的權重，通常設置的方法為negative samples / positive samples，也就是少數的權重會較高，多數的會較低，所以模型在訓練的時候就會比較傾向於多預測少數的樣本，而就在調整該項參數之後，模型的表現就直逼排行榜的第一梯隊。

* + Voting

當我們發現不論再怎麼調參或是多做data preprocess都不太能再增進一點的時候，我們想到Ensemble的觀念：若是我們將Boosting系列的模型結合起來呢？因為上述三者在實作上的細節還是各不相同，透過feature importance也可以看出三者對於不同feature的關注度是不一樣的，那麼若是我們可以結合三個模型共同去做預測，那麼可能某個模型預測不好的樣本就可以被其他模型所糾正，從而再提升一點performance，所以我們使用VotingClassifier將三者結合起來再做共同預測。

* + Neural Network

在課程教到Neural Network的時候，我們也有嘗試實作類似模型。首先嘗試的是簡單的MLP，僅由三層fully connected layers所組成，而隱藏層套用的activation為relu，輸出層則為sigmoid。雖然在validation set的F1-score表現都不錯，但在Kaggle上的分數卻不高，在增加層數及調整參數過後得到的結果也差不多。

隨後我們有嘗試RNN的類型，例如LSTM，但是可能是因為資料間不具有連續性，所以該類模型的表現效果更差。

* + K-Means

我們也有試過Clustering的方法，K-Means即為較具代表性的，其原理基本上就是對於每個data去計算其鄰近K個data為何，透過他們的label去predict該data的結果，也就是一種分群的概念。而在訓練過程中，由於只有0和1，所以我們思考clusters為2的情況應該會有不錯的成果，但似乎是受限於資料不平衡或者是資料間關聯不是很大的情況，得到的score反而是三者中最差