Model

建立模型：

這部分為股票價格預測模型的建模。主要建的模型目的是預測未來五日每一支股票平均收盤是否高於今日，以此提供我們是否進出場的依據之一。

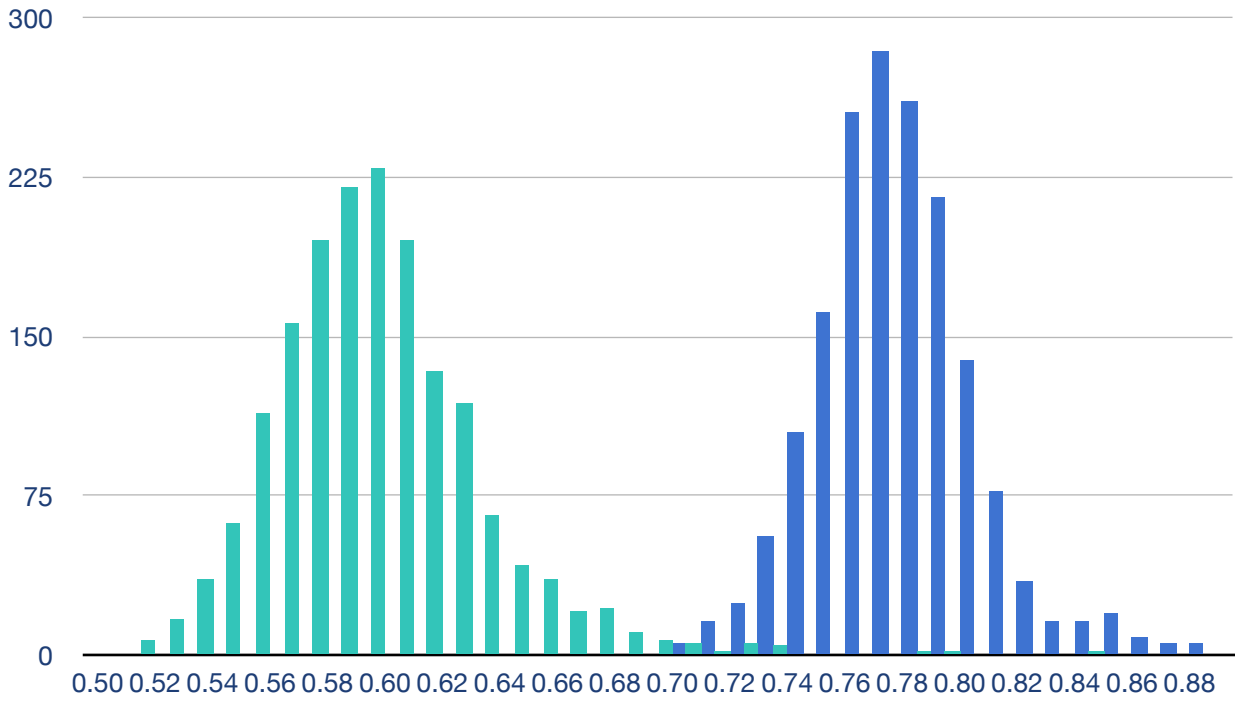
特徵選擇：

首先模型最重要的部分，就是特徵的選擇，我們原始數據有每日的開高收低量，從這五個特徵，我們可以利用這五個基本變量，近而衍生出其他股市的技術性指標像是macd10、rsv9…等，透過變數的選擇，最終我們選擇了包括開高收低量等十五個變量來做為我們後續模型建模的基礎。

原始數據：Open 、High、Low、Close、Volume

技術指標：MACD_10、RSV_9、MA_5、Mean_Volume . . . . . .

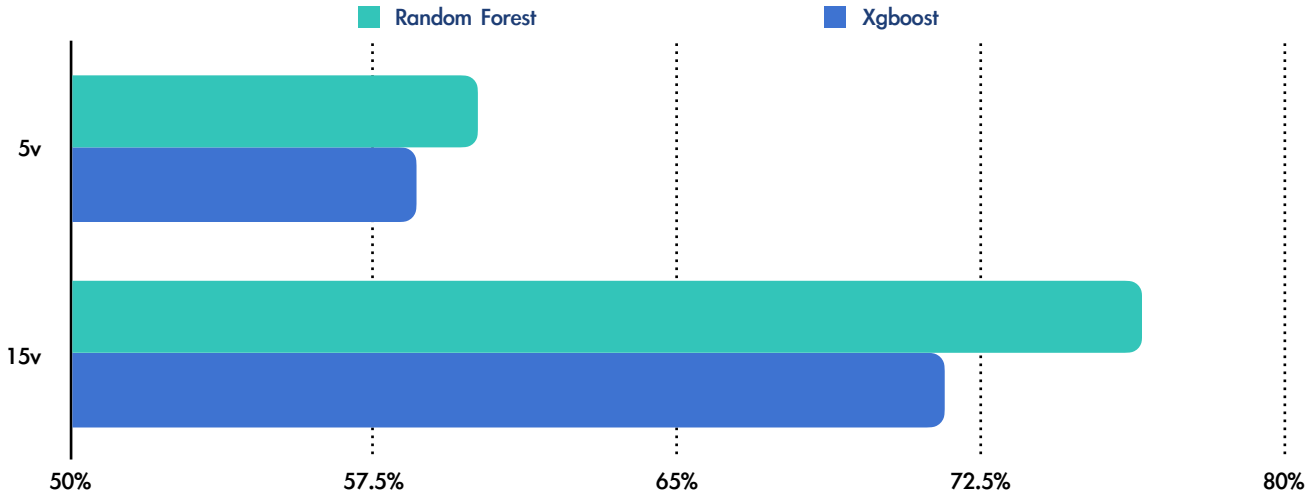
一開始我們選用隨機森林進行建模，挑選的特徵為原始資料的每日開高收低量，但是做出的結果其準確率約落在60%上下。當我們選用原始資料特徵以及技術性指標，準確率將由原本的60.1%提升至到76.5%左右。



所以特徵的選擇真的很重要。但是我們還是想進一步的提高整體預測的準確度，因此此時出現了兩個方向，其一為，選擇更適合的預測模型，其二為對模型進行Tunning。

其他模型：

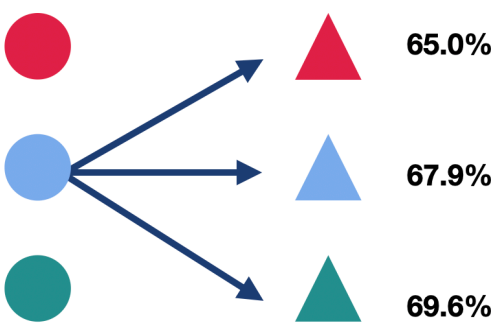
這邊我們先以選用其他模型為第一考量，我們選用的模型是最近很紅的Xgboost。但做出的結果卻大大出乎意料。發現隨機森林的表現不管是在五個變數下，亦或是十五個變數之下的表現結果有顯著的比Xgboost所得到的準確率高。

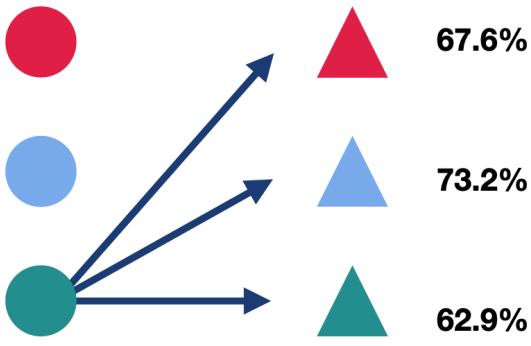


因此我們合理的猜測，隨機森林模型在小樣本的數據量前提下，有著比起Xgboost有著較好的使用效果。
做到這裡，我們也先暫時選擇放棄使用其他模型這個方向，轉由選擇優化原本的模型。

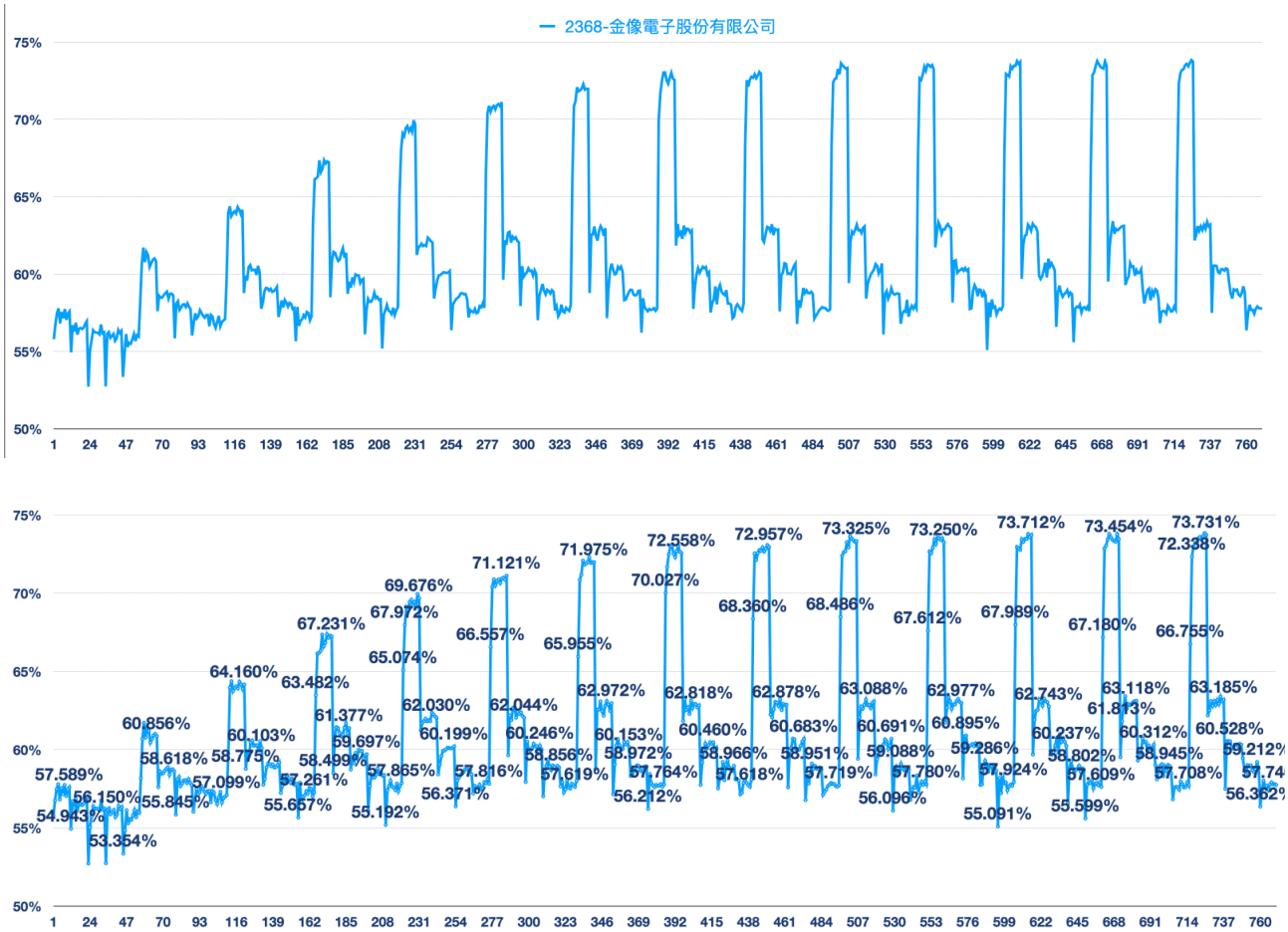
Tunning(調參)：

這邊先來說明一下什麼是tuning。假設一個模型共有兩個超參數，分別以圓形及三角形來形象化。接著分別有紅藍綠三種組合，我們依次將他們配對，即可獲得相應的準確率，我們可以看到率圈圈配上藍色三角形準確率最高，所以圈圈超參數就選用綠色，三角形超參數選用藍色。





於是就把其中一支股票拿出來看看，x軸為組合代號，y軸為準確率，我們可以發現有四個很類似的波型，是因為我們只針對四個超參數進行tuning。



我們可以看使用預設模型，跟tuning完的結果會有顯著的差異。

雖然以結果論來看，進步的幅度看起來不大，那是因為我們是以原本準確率就不高的股票來進行調參，因此實際有做tuning的股票其實只有了了幾支而已。

我們總共有1720支上市上櫃的股票，換句話說，我們有1720支不同的股票模型。
這邊大家可能會有一個問題，就是為什麼我們不對每支股票都進行tuning呢?原因有兩點，其一為時間成本的考量，其二為機器的原因，因此我們才只針對準確率低於70%的進行調參。但是可以預期的是如果未來我們可以對每一支股票模型都進行調參的話，那勢必可大幅度的提升我們模型的平均準確率。
那我們如何能以最低的成本來再次優化我們的模型呢?

Before Tuning	After Tuning
76.5%	76.6%

Stacking：

此時我們回頭想到了之前在選擇模型時所做的Xgboost模型，於是結合Stacking的精神，將Xgboost模型與隨機模型結合，以AUC 準確率做簡單的權重分配，讓程式內部替我們決定要使用哪一個模型進行預測。當經過Stacking後，即可發現我們總體的股票模型平均準確率又再次的提升。

Before Stacking	After Stacking
76.6%	76.7%

主講人：周盟鎮

https://github.com/MengChenChou