Random Forest

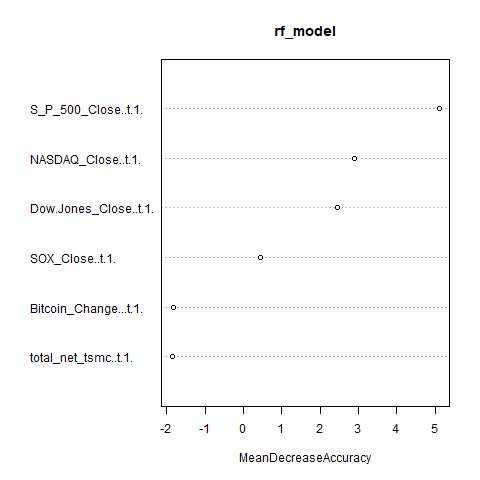
(1)dataset1: ourdata.csv

本模型將ourdata保留最後100筆當作unseen的testing data，其他資料以課程作業方式分成training, validating及testing data進行訓練與評估，其中null model是以預測全部為1來進行，並將結果寫入performance.csv

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AUC | Training | Validation | Testing | Accuracy | Training.1 | Validation.1 | Testing.1 | Precision | Training.2 | Validation.2 | Testing.2 | Recall | Training.3 | Validation.3 | Testing.3 |
| fold1 | 1 | 0.86 | 0.88 | fold1 | 0.98 | 0.79 | 0.78 | fold1 | 0.98 | 0.79 | 0.77 | fold1 | 0.98 | 0.87 | 0.85 |
| fold2 | 1 | 0.87 | 0.88 | fold2 | 0.99 | 0.81 | 0.79 | fold2 | 0.99 | 0.81 | 0.81 | fold2 | 0.99 | 0.85 | 0.84 |
| fold3 | 1 | 0.9 | 0.86 | fold3 | 0.98 | 0.82 | 0.79 | fold3 | 0.97 | 0.8 | 0.78 | fold3 | 0.99 | 0.88 | 0.85 |
| fold4 | 1 | 0.89 | 0.88 | fold4 | 0.99 | 0.8 | 0.82 | fold4 | 0.99 | 0.81 | 0.78 | fold4 | 0.99 | 0.84 | 0.9 |
| fold5 | 1 | 0.92 | 0.86 | fold5 | 0.98 | 0.85 | 0.79 | fold5 | 0.98 | 0.84 | 0.79 | fold5 | 0.99 | 0.9 | 0.84 |
| ave. | 1 | 0.89 | 0.87 | ave. | 0.98 | 0.82 | 0.79 | ave. | 0.98 | 0.81 | 0.79 | ave. | 0.99 | 0.87 | 0.85 |
| AUC | null\_Training | null\_Validation | null\_Testing | Accuracy | null\_Training.1 | null\_Validation.1 | null\_Testing.1 | Precision | null\_Training.2 | null\_Validation.2 | null\_Testing.2 | Recall | null\_Training.3 | null\_Validation.3 | null\_Testing.3 |
| fold1 | 0.5 | 0.5 | 0.5 | fold1 | 0.55 | 0.58 | 0.54 | fold1 | 0.55 | 0.58 | 0.54 | fold1 | 1 | 1 | 1 |
| fold2 | 0.5 | 0.5 | 0.5 | fold2 | 0.55 | 0.54 | 0.58 | fold2 | 0.55 | 0.54 | 0.58 | fold2 | 1 | 1 | 1 |
| fold3 | 0.5 | 0.5 | 0.5 | fold3 | 0.54 | 0.52 | 0.54 | fold3 | 0.54 | 0.52 | 0.54 | fold3 | 1 | 1 | 1 |
| fold4 | 0.5 | 0.5 | 0.5 | fold4 | 0.54 | 0.55 | 0.52 | fold4 | 0.54 | 0.55 | 0.52 | fold4 | 1 | 1 | 1 |
| fold5 | 0.5 | 0.5 | 0.5 | fold5 | 0.55 | 0.54 | 0.55 | fold5 | 0.55 | 0.54 | 0.55 | fold5 | 1 | 1 | 1 |
| ave. | 0.5 | 0.5 | 0.5 | ave. | 0.55 | 0.55 | 0.55 | ave. | 0.55 | 0.55 | 0.55 | ave. | 1 | 1 | 1 |

評估後將unseen以外的資料再進行模型訓練，並以unseen的100筆資料進行測試，發現accuracy僅約0.55，此與前述的評估差異甚大

另進行varImp (very important variable)測試及繪圖，考量參數有限，因此著手將現有dataset資料進行轉換，以增加features，包括所有指標的漲或跌(1,0)、4大指標取log值、台股前一日漲或跌，此外並將新增指標漲跌合計(其中S&P500賦予權重6、台股3、其他1 )



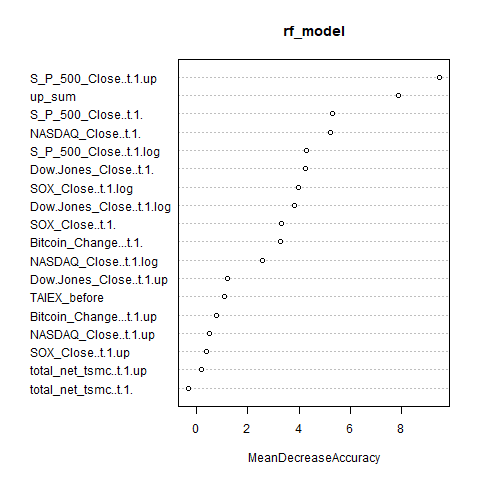
(2)dataset1: ourdata\_addFeatures.csv

將ourdata\_addFeatures保留最後100筆當作最後unseen的testing data，其他資料以課程作業方式分成training, validating及testing data進行訓練與評估，其中null model是以預測全部為1來進行，並將結果寫入performance.csv

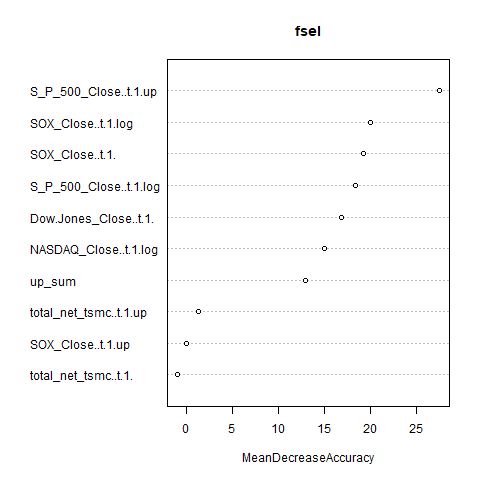
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AUC | Training | Validation | Testing | Accuracy | Training.1 | Validation.1 | Testing.1 | Precision | Training.2 | Validation.2 | Testing.2 | Recall | Training.3 | Validation.3 | Testing.3 |
| fold1 | 0.99 | 0.87 | 0.82 | fold1 | 0.95 | 0.8 | 0.76 | fold1 | 0.95 | 0.79 | 0.79 | fold1 | 0.96 | 0.88 | 0.78 |
| fold2 | 0.99 | 0.88 | 0.89 | fold2 | 0.96 | 0.82 | 0.82 | fold2 | 0.97 | 0.86 | 0.85 | fold2 | 0.95 | 0.8 | 0.83 |
| fold3 | 0.99 | 0.88 | 0.89 | fold3 | 0.96 | 0.81 | 0.8 | fold3 | 0.96 | 0.84 | 0.84 | fold3 | 0.96 | 0.76 | 0.8 |
| fold4 | 1 | 0.89 | 0.89 | fold4 | 0.96 | 0.86 | 0.82 | fold4 | 0.96 | 0.86 | 0.79 | fold4 | 0.98 | 0.88 | 0.89 |
| fold5 | 0.99 | 0.85 | 0.88 | fold5 | 0.95 | 0.8 | 0.82 | fold5 | 0.93 | 0.79 | 0.8 | fold5 | 0.98 | 0.86 | 0.88 |
| ave. | 0.99 | 0.88 | 0.87 | ave. | 0.96 | 0.81 | 0.81 | ave. | 0.95 | 0.83 | 0.81 | ave. | 0.96 | 0.84 | 0.83 |
| AUC | null\_Training | null\_Validation | null\_Testing | Accuracy | null\_Training.1 | null\_Validation.1 | null\_Testing.1 | Precision | null\_Training.2 | null\_Validation.2 | null\_Testing.2 | Recall | null\_Training.3 | null\_Validation.3 | null\_Testing.3 |
| fold1 | 0.5 | 0.5 | 0.5 | fold1 | 0.55 | 0.56 | 0.56 | fold1 | 0.55 | 0.56 | 0.56 | fold1 | 1 | 1 | 1 |
| fold2 | 0.5 | 0.5 | 0.5 | fold2 | 0.52 | 0.56 | 0.56 | fold2 | 0.52 | 0.56 | 0.56 | fold2 | 1 | 1 | 1 |
| fold3 | 0.5 | 0.5 | 0.5 | fold3 | 0.54 | 0.51 | 0.56 | fold3 | 0.54 | 0.51 | 0.56 | fold3 | 1 | 1 | 1 |
| fold4 | 0.5 | 0.5 | 0.5 | fold4 | 0.56 | 0.53 | 0.51 | fold4 | 0.56 | 0.53 | 0.51 | fold4 | 1 | 1 | 1 |
| fold5 | 0.5 | 0.5 | 0.5 | fold5 | 0.56 | 0.56 | 0.53 | fold5 | 0.56 | 0.56 | 0.53 | fold5 | 1 | 1 | 1 |
| ave. | 0.5 | 0.5 | 0.5 | ave. | 0.55 | 0.54 | 0.54 | ave. | 0.55 | 0.54 | 0.54 | ave. | 1 | 1 | 1 |

接著將unseen以外的資料進行模型訓練，並以unseen的100筆資料進行測試，發現accuracy約為約0.53，仍低於之前評估資料，且單純以「S&P500漲或跌」當作預測結果，其accuracy可達0.64

接著進行varImp (very important variable)測試及繪圖



透過varImp進行參數選擇如下



最後再經tuning，選擇S&P500漲跌、台積電(total\_net\_tsmc..t.1.)、彼特幣(Bitcoin\_Change...t.1.)、漲跌權重相加(up\_sum)、費城半導體取log(SOX\_Close..t.1.log)及台股前一日漲跌(TAIEX\_before)當作變數進行訓練，accuracy提升至約0.62

