

Lab05_110062401_Report

1. 目錄

- a. Dataset
- b. 用了哪一些Model以及對應的準確率等表現
- c. 做了哪一些調整，改變，以及觀察到結果
- d. 如何選取 Top 3 important features
- e. 評估模型的Error Curve說明是否Overfit或者Underfit

2. Dataset

- a. x_train (68600, 8)

	PERIOD	GAME_CLOCK	SHOT_CLOCK	DRIBBLES	TOUCH_TIME	SHOT_DIST	PTS_TYPE	CLOSE_DEF_DIST
0	1	358	2.4	0	3.2	20.6	2	4.5
1	1	585	8.3	0	1.2	3.0	2	0.5
2	1	540	19.9	0	0.6	3.5	2	3.2
3	1	392	9.0	0	0.9	21.1	2	4.9
4	3	401	22.7	0	0.7	4.1	2	2.9

- b. x_test (17151, 8)

	PERIOD	GAME_CLOCK	SHOT_CLOCK	DRIBBLES	TOUCH_TIME	SHOT_DIST	PTS_TYPE	CLOSE_DEF_DIST
0	3	595	11.3	1	1.8	3.9	2	0.3
1	2	530	11.0	0	1.0	24.3	3	6.3
2	1	221	21.3	2	1.9	3.5	2	11.5
3	3	442	9.0	0	0.6	2.4	2	3.4
4	1	634	16.1	0	0.8	4.2	2	1.6

3. 用了哪一些模型以及其表現

a. RandomForestClassifier

根據assignment需求，選擇了之前lab學過的RandomForestClassifier來當作這次的模型。在還未做Hyperparameters調整，以及選取features前，單純模型的執行結果為如下，看起來還蠻不錯了。

```
===== RandomForestClassifier =====
MSE train: 0.36, valid: 0.38
Acc train: 0.64, valid: 0.62
=====
```

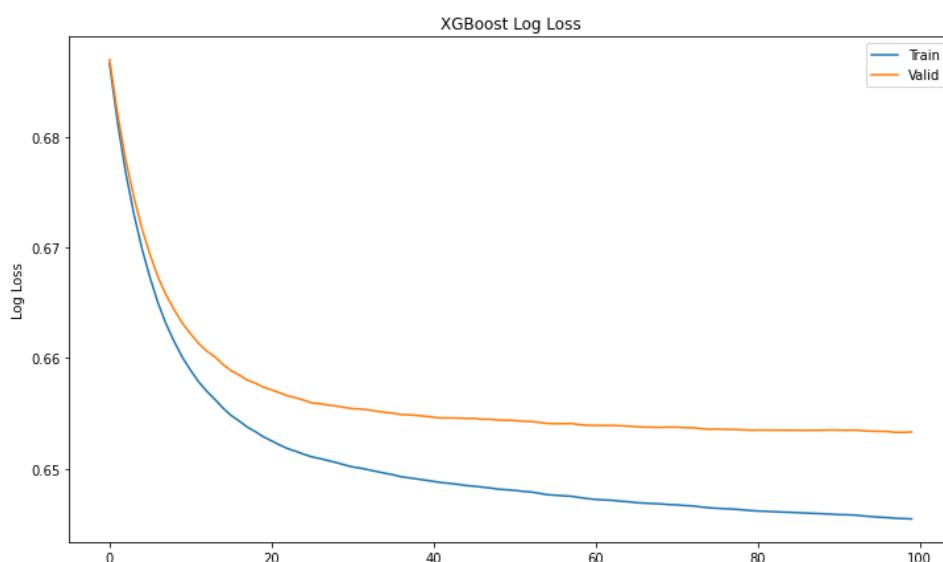
valid_acc = 0.62

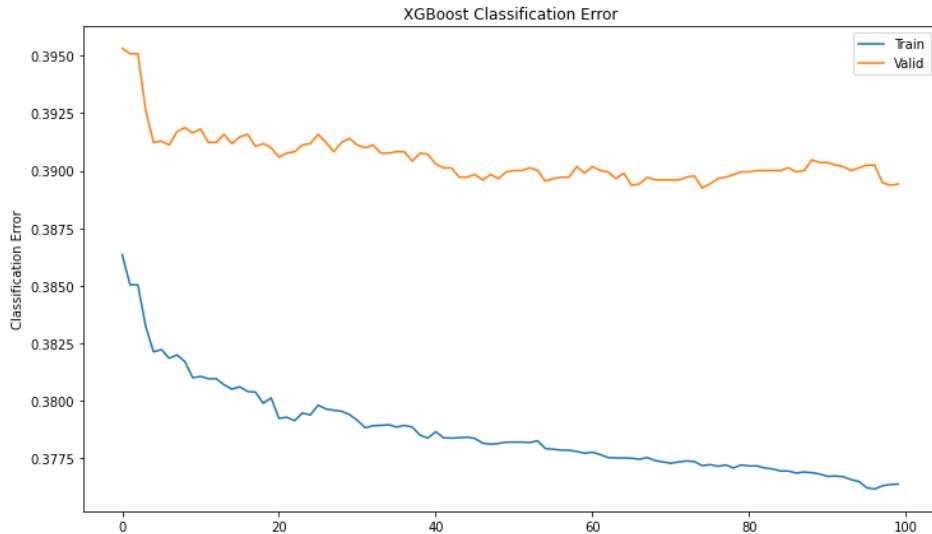
b. XGBoost

```
===== XGBoost =====
MSE train: 0.38, valid: 0.39
Acc train: 0.62, valid: 0.61
=====
```

除了課堂教的模型外，這裡還額外選擇了XGBoost作為對比，objective為binary:logistic，也是跑出來約0.62的結果，證明該dataset給與的資訊最多也只能讓模型達到6成的準確率效果。

valid_acc = 0.61





4. 做了哪一些調整，改變，以及觀察到結果

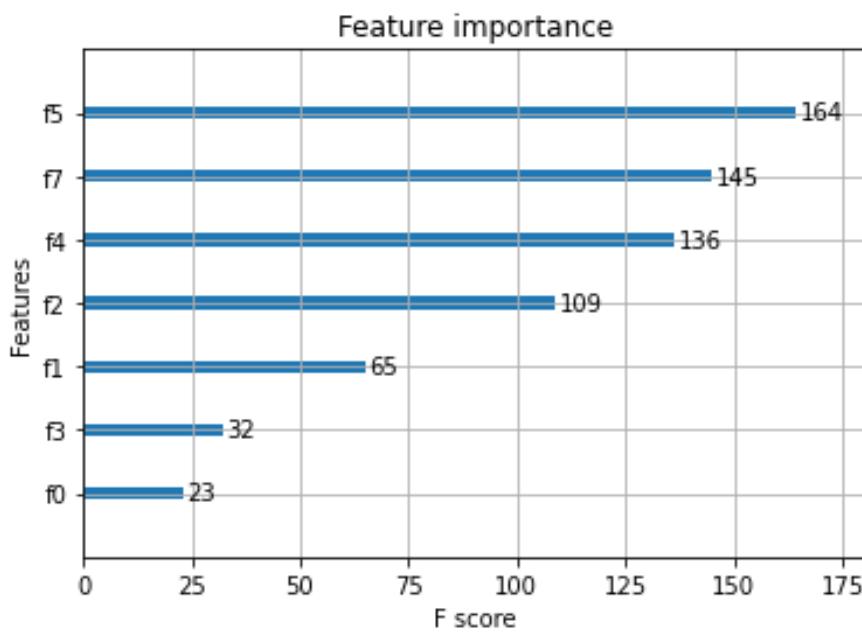
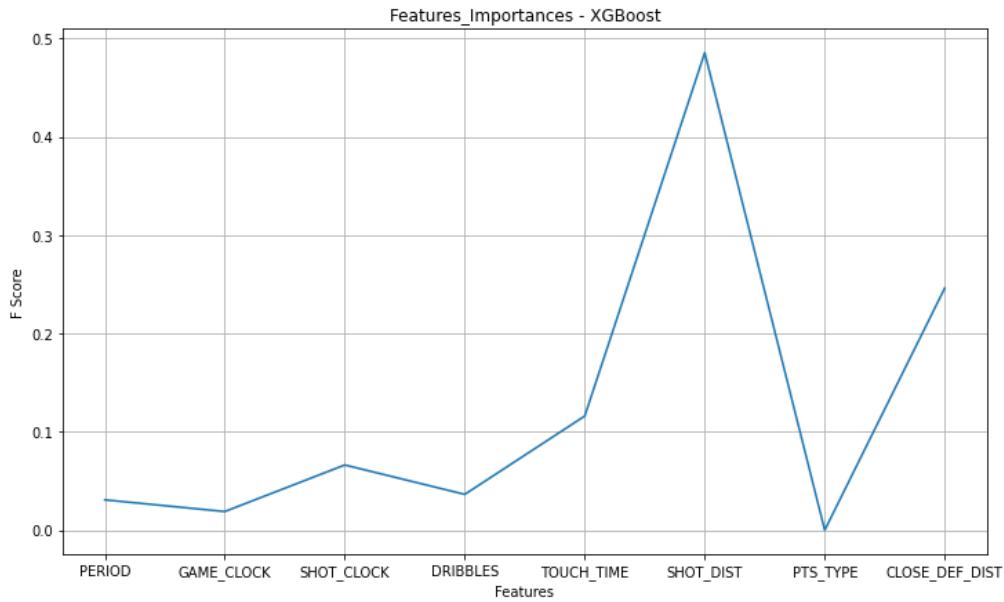
- 調整過train 和 valid dataset 的比例(7:3, 8:2), 但似乎結果都沒有太大差別，都是落在0.61~0.62左右。
- 通過增加模型樹的深度，來使模型複雜度提高，是可以使得模型的準確率上升，但是valid_mse下降的值最大只有約為0.02, valid_acc上升0.02
 -
- 以下附上

5. 如何選取 Top 3 important features

- 通過課堂上教過的Lasso, 設定alpha=0.01, epsilon=1e-2, 可以得到四個features，分別是 'SHOT_CLOCK', 'TOUCH_TIME', 'SHOT_DIST', 'CLOSE_DEF_DIST'。

```
Selected attributes: ['SHOT_CLOCK' 'TOUCH_TIME' 'SHOT_DIST' 'CLOSE_DEF_DIST']
```

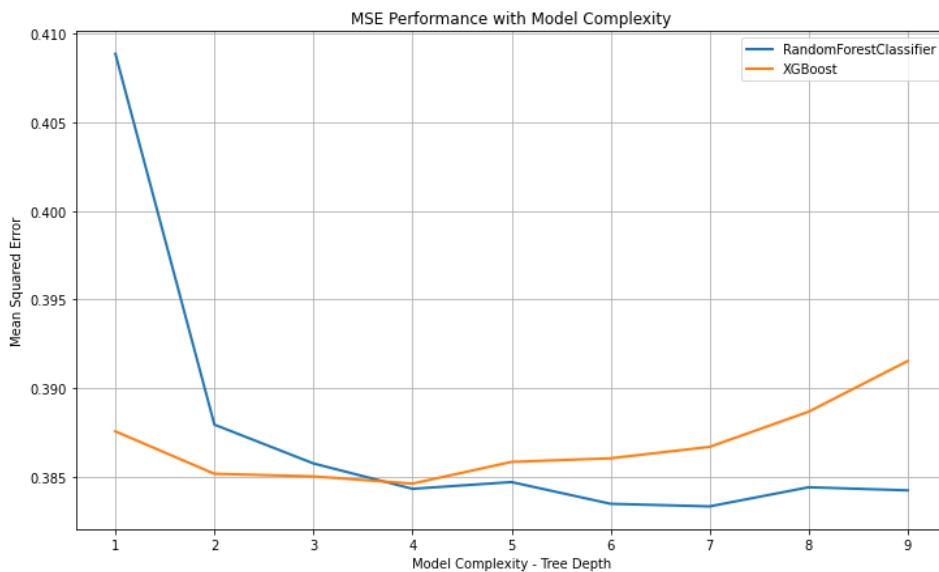
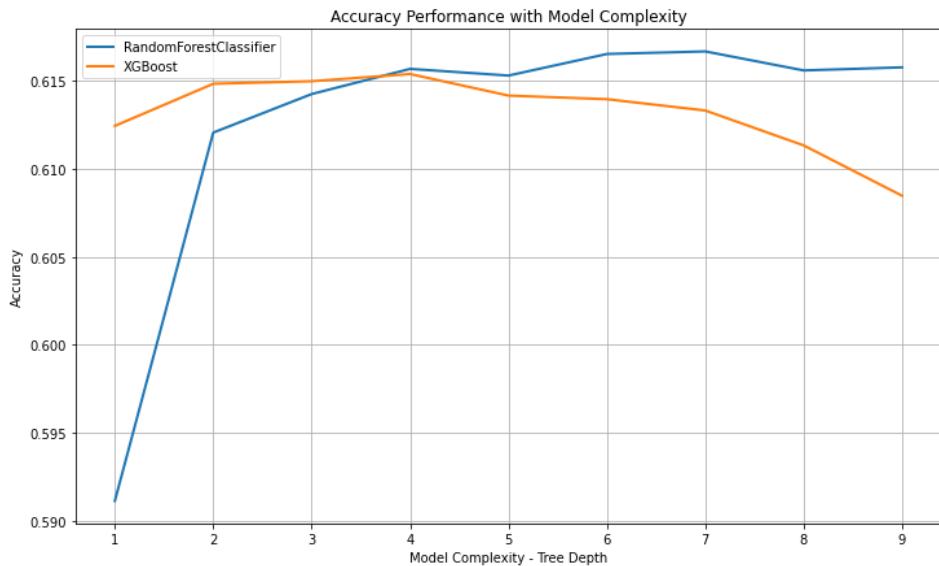
- 再來是通過XGBoost模型defined的API, features_important_會顯示出模型中的features權重比值，可以從中看出哪一些features比較重要。



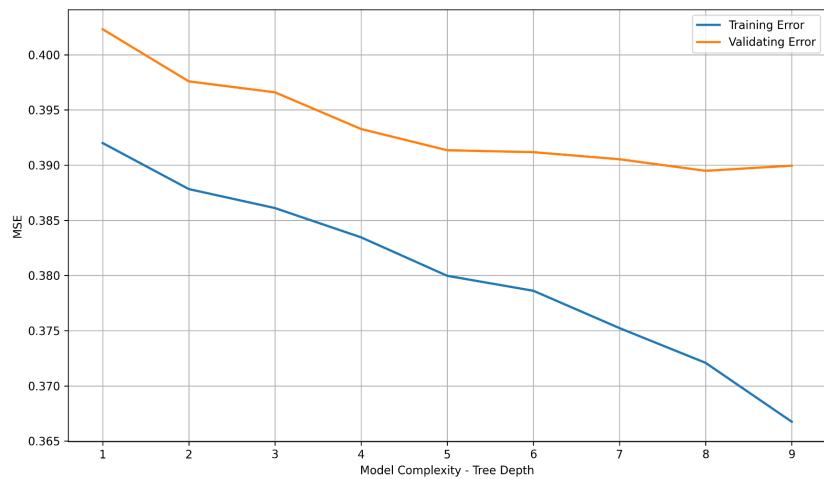
- 因此我選擇了 **SHOT_DIST, CLOSE_DEF_DIST, TOUCH_TIME** 這三個features為主

6. 評估 Error Curve 是否有Overfit 或者 Underfit

- 以下的圖顯示以上提到的兩個模型對Valid dataset在隨著模型的複雜度提高時，準確率也會提升，但是對於RFC來說最好的深度是Depth 7可以使得準確率最高，對於XGB最好的準確率是在Depth 4。
- loss的話，RFC是Depth為7最低，XGB是Depth為4最低，和Acc的表現相呼應。



- 最後選表現最好的RFC單獨來跑depth，選出MSE最低的Depth，預測x_test的target values。



- RFC的Validate Error最低是Depth 8
- **x_test Prediction:**

	A
1	FGM
2	1
3	0
4	1
5	1
6	1
7	0
8	0
9	0
10	0