

期末專題

Kaggle's Spaceship Titanic

110321006 蔡秉翰

110321007 劉竑毅

1. 介紹

泰坦尼克號宇宙飛船是一個月前下水的星際客輪。這艘船上有近 **13,000** 名乘客，將來自太陽系的移民運送到三顆新近居住的系外行星，這些系外行星圍繞著附近的恆星運行。

在繞過半人馬座阿爾法星前往其第一個目的地 - 炎熱的 **55 Cancri E** 時，粗心的泰坦尼克號宇宙飛船與隱藏在塵埃雲中的時空異常相撞。可悲的是，它遇到了與 **1000** 年前同名的名字相似的命運。雖然這艘船完好無損，但幾乎一半的乘客被運送到另一個維度！

2. 數據準備和預處理

數據讀取： 從"**train.csv**"和"**test.csv**"文件中讀取訓練和測試數據。

缺失值處理： 通過填充類別型和數值型特徵的缺失值，確保數據完整性。

PassengerId	0.000	PassengerId	0
HomePlanet	2.221	HomePlanet	0
CryoSleep	2.390	CryoSleep	0
Cabin	2.305	Cabin	0
Destination	2.113	Destination	0
Age	2.082	Age	0
VIP	2.282	VIP	0
RoomService	2.028	RoomService	0
FoodCourt	2.228	FoodCourt	0
ShoppingMall	2.359	ShoppingMall	0
Spa	2.190	Spa	0
VRDeck	2.066	VRDeck	0
Name	2.267	Name	0

將資料缺失的部分填入平均值

```
data_1 = clipping_quantile(data_1, None, 0.99)
```

移除極端值

3. 特徵工程

特徵提取： 根據數據中的不同特徵創建新的特徵。

PassengerGroup 特徵：

使用 **PassengerId** 列的下劃線分割，提取分組標籤，
並將其存儲在新的 **PassengerGroup** 列中。

IsAlone 特徵：

通過對 **PassengerGroup** 進行分組，計算每個分組中的乘客數量，並創建一個新的特徵 **IsAlone**，表示是

否獨自一人。如果乘客數量大於 1，則標記為"Not Alone"，否則標記為"Alone"。

CabinDeck、DeckPosition、CabinSide 特徵：

根據 Cabin 列的信息，創建三個新的特徵。

CabinDeck 提取 Cabin 的首個字母。

DeckPosition 根據 CabinDeck 的值，將艙位分為 "Lower"或"Higher"。

CabinSide 提取 Cabin 中斜槓後的部分。

Regular、Luxury、TotalSpendings 特徵：

基於房間服務、餐廳和購物等支出類別，創建三個新的特徵。

Wealthiest_Deck 特徵：

根據 CabinDeck 分組，計算每個艙位的總支出和乘客數量。

創建 DeckAverageSpent 特徵，表示每個艙位的平均支出。

FamilyName、NoRelatives、FamilySizeCat 特徵：

從 Name 中提取家庭名稱，並基於家庭成員數量創建相應的分組標籤。

類別型特徵編碼： One-Hot Encoding 將類別型特徵轉換為可供模型訓練的形式。

4. 模型建構與優化

RandomForestClassifier

使用隨機森林進行模型建構。

```
0.7867290980516646
```

通過網格搜索優化超參數，提高模型性能。

```
Best Hyperparameters: {'max_depth': 7, 'n_estimators': 61}
```

```
Mean Accuracy: 0.802953811128767
```

GradientBoostingClassifier

利用梯度提升樹進行模型建構。

```
0.8005378992531383
```

通過網格搜索優化超參數，提高模型性能。

```
Best Hyperparameters: {'max_depth': 5, 'n_estimators': 81}
```

```
Mean Accuracy: 0.8023777742465172
```

XGBClassifier

使用 XGBoost 進行模型建構。

```
0.7973120928015255
```

通過網格搜索優化超參數，提高模型性能。

```
Mean Accuracy: 0.8078963398485091
```

```
Best Hyperparameters: {'max_depth': 5, 'n_estimators': 41}
```

5. 模型集成

VotingClassifier：將優化過的

RandomForestClassifier、GradientBoostingClassifier 和

XGBClassifier 進行投票集成，提高整體預測性能。

voting=soft

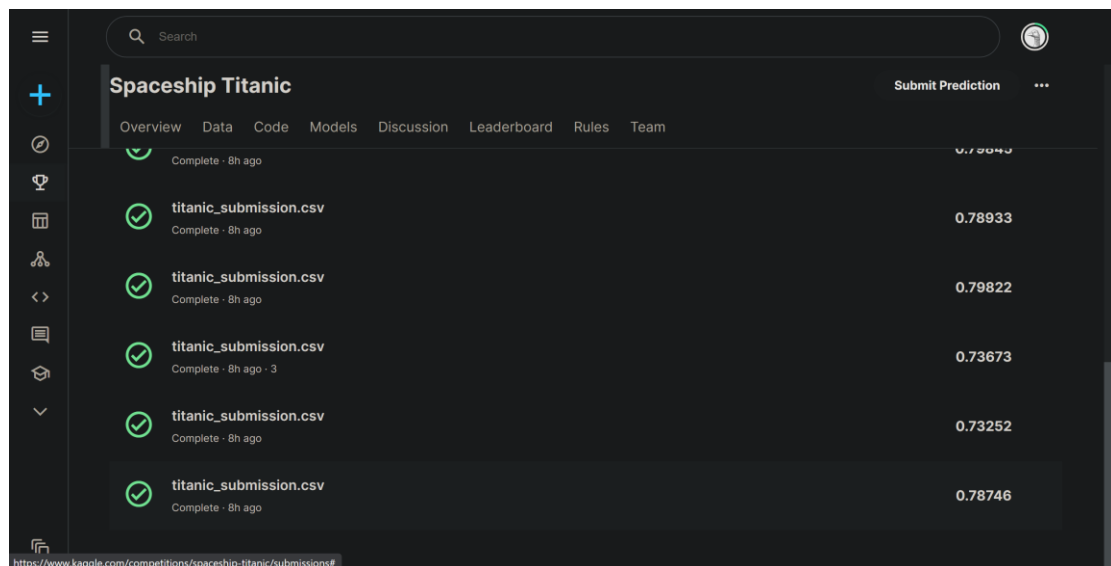
```
Mean Accuracy: 0.8046724073925798
```

6. 模型評估

將 RandomForestClassifier、GradientBoostingClassifier

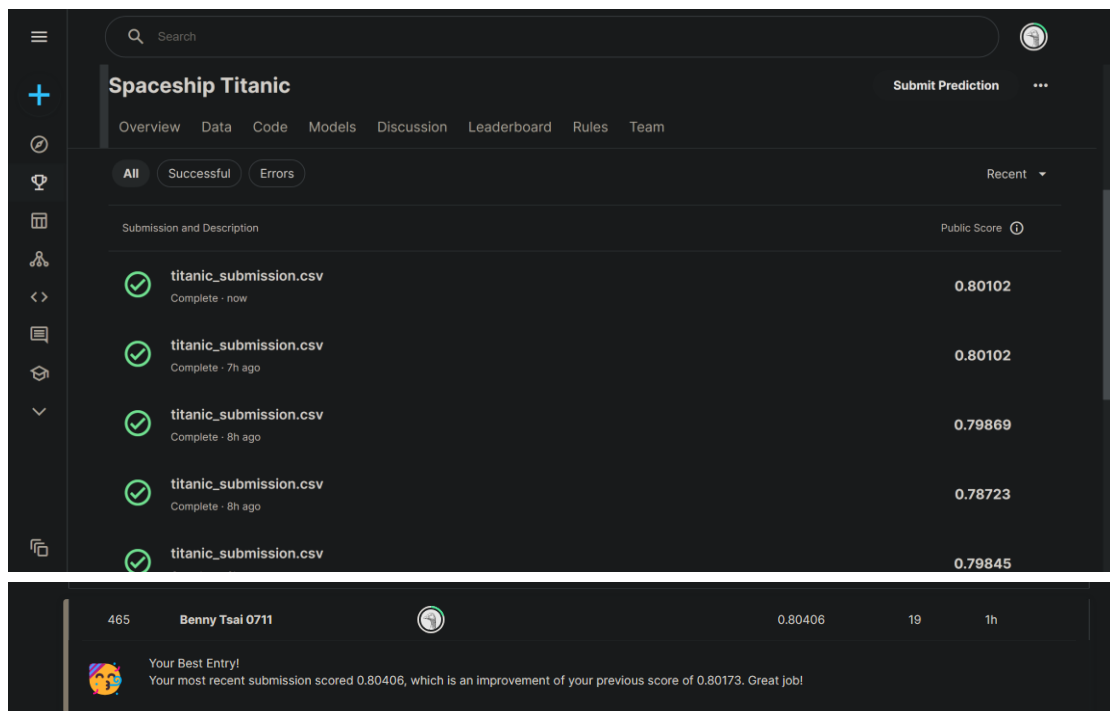
和 XGBClassifier、VotingClassifier 都上傳評分得到準確

率最高的是 VotingClassifier



The screenshot shows the Kaggle Spaceship Titanic competition page. The 'Leaderboard' tab is selected, displaying a list of submissions. Each submission is represented by a green checkmark icon, the filename 'titanic_submission.csv', a status 'Complete - 8h ago', and a score. The scores are listed in descending order: 0.78943, 0.78933, 0.79822, 0.73673, 0.73252, and 0.78746. The URL at the bottom is <https://www.kaggle.com/competitions/spaceship-titanic/submissions#>.

Submission	Score
titanic_submission.csv (Complete - 8h ago)	0.78943
titanic_submission.csv (Complete - 8h ago)	0.78933
titanic_submission.csv (Complete - 8h ago)	0.79822
titanic_submission.csv (Complete - 8h ago - 3)	0.73673
titanic_submission.csv (Complete - 8h ago)	0.73252
titanic_submission.csv (Complete - 8h ago)	0.78746



經過反覆測試之後得到最終結果

準確率 0.80406

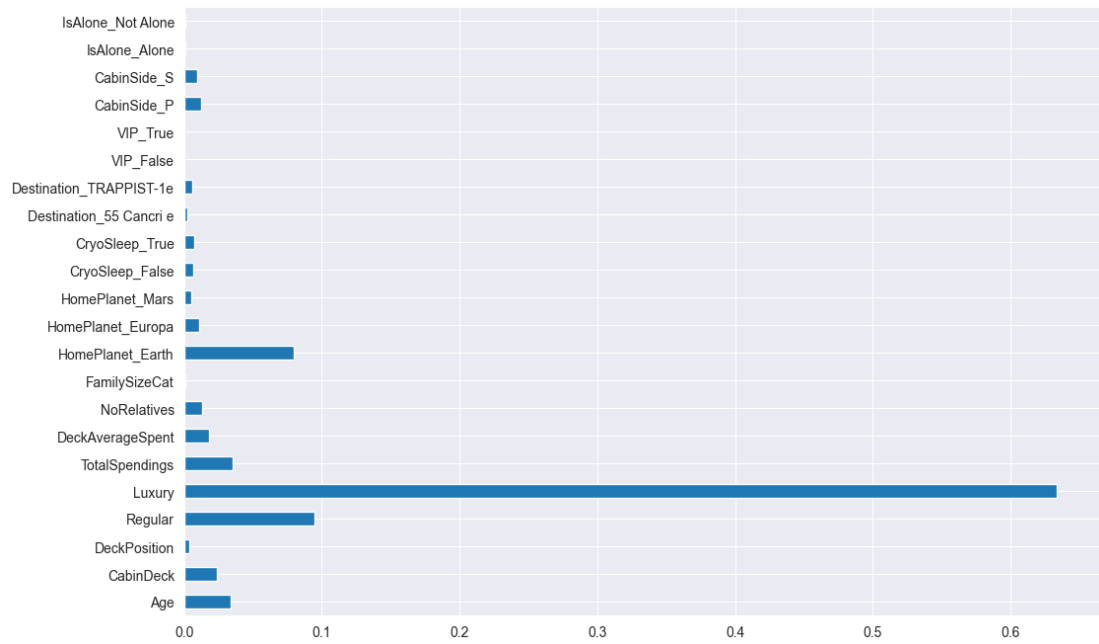
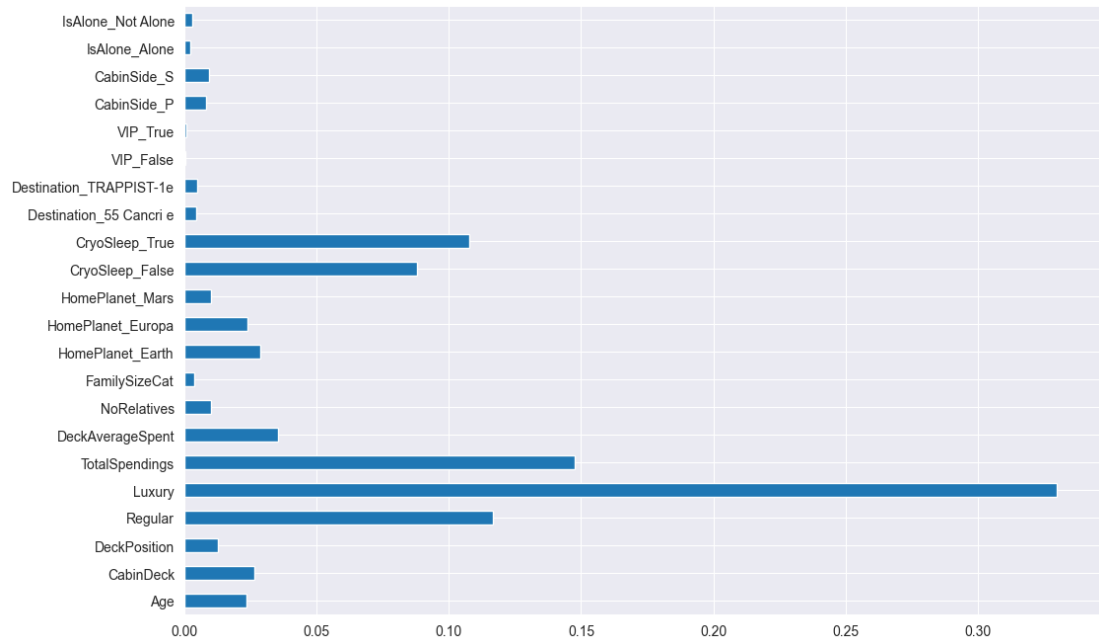
排名 465

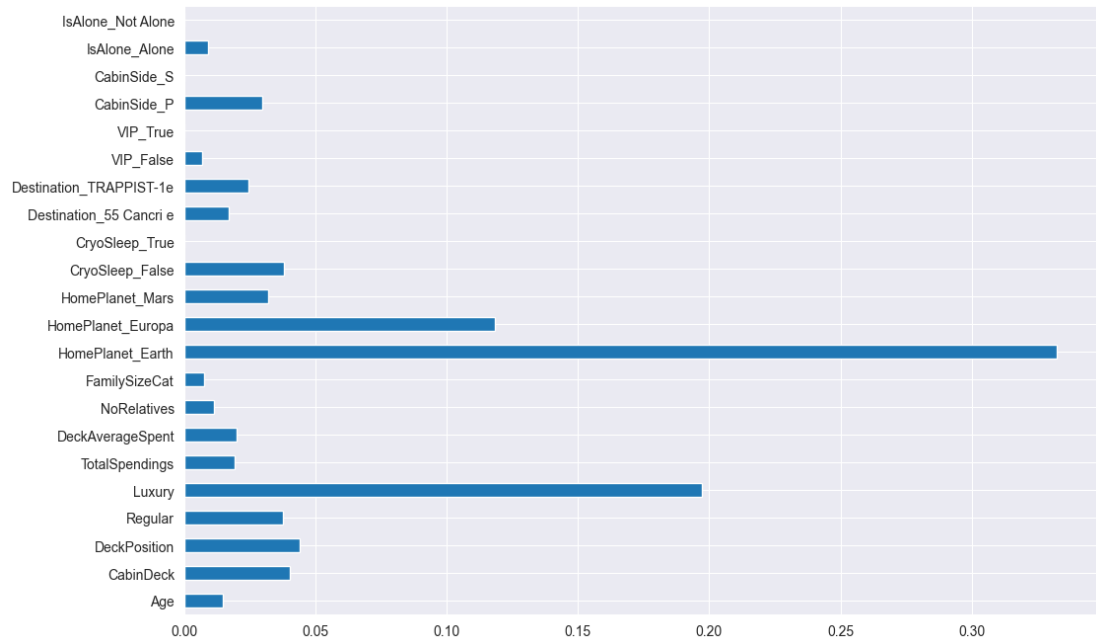
全部人數 2890

前百分之 16

7. 特徵重要性

特徵重要性可視化： 通過繪製每個模型的特徵重要性圖，了解模型對特徵的重要性排名。





由上到下分別是 RandomForestClassifier 、

GradientBoostingClassifier 和 XGBClassifier

可以看出 GB,XGB 有正則化所以有些特徵會歸零

未來可能可以透過特徵篩選提高準確率