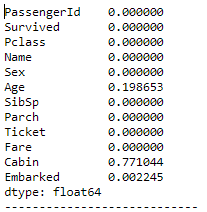
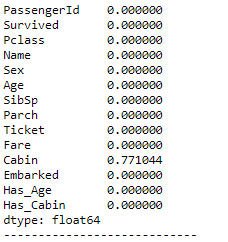
**310706034 資管碩一 吳啓玄 資料探勘研究與實務 HW0- titanic小練習**

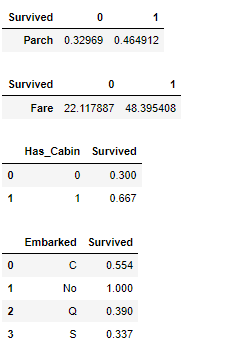
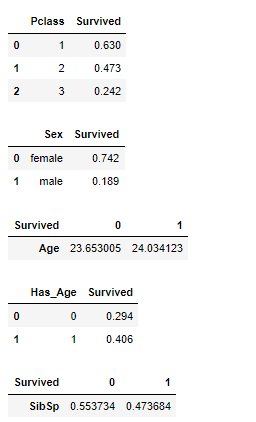
1. 觀察下圖訓練集資料變數遺失值比例，發現Age、Cabin、Embarked有遺失值，其中Age、Cabin有較大比例遺失值，新增Has\_Age、Has\_Cabin變數，若Age、Cabin是遺失值，則Has\_Age、Has\_Cabin為0，反之為1，另外，Embarked補值 ”No”，Age補值0。



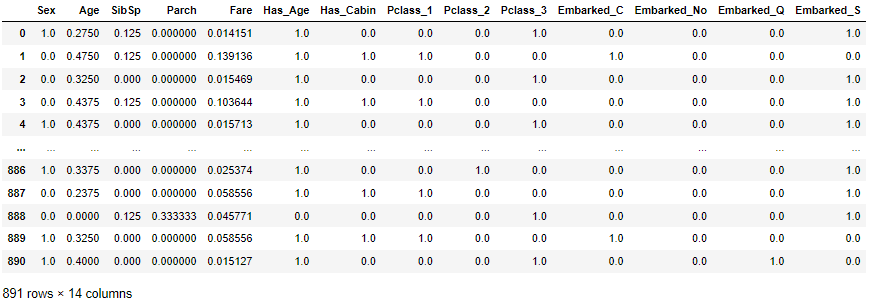
處理後之訓練集資料變數遺失值比例如下:



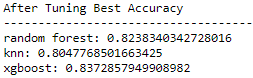
1. 針對有興趣之變數進行資料探索性分析(EDA)，觀察變數存活率，如下:



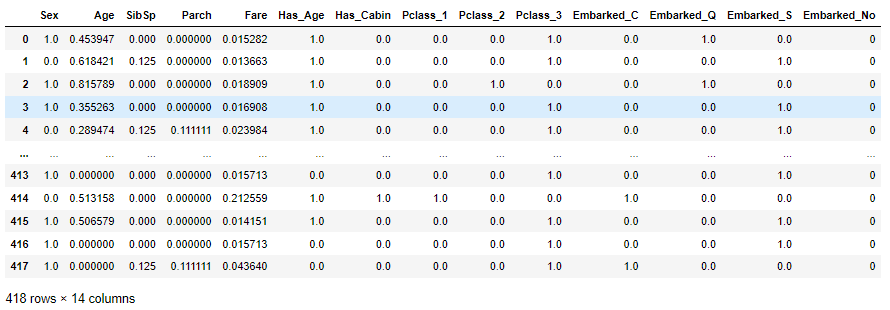
1. 把訓練集資料轉換為dummy variable並做min-MAX scale，轉換後如下:



1. 建構Random Forest、XGBoost、KNN三個模型，根據5-fold Cross Validation進行tuning，Random Forest調參樹木數量、樹木深度，XGBoost調參樹木數量、樹木深度，KNN調參觀看鄰居數，得到以下結果並挑選準確率最高模型



1. 測試集資料沒有Embarked=”No”所以補一欄，另外，Fare有遺失值所以採取平均插值，其餘皆和訓練集資料處理方式相同，測試集資料處理後如下:



1. 使用XGBoost模型，對所有訓練集資料依照cross validation tuning得到的最佳參數進行建模，最後，用該模型對測試集做預測後，上傳Kaggle的準確率是0.77033: 