**資料探勘研究與實務**

0853426 陳紹雲

學生首先載入資料,將資料重組為題目給訂的 feature&label, 爾後將欄位為" NONE" 的值改為 0, 在將 features_dummy = ['DayOfWeek', 'PdDistrict', 'Dates_Hours'] 使用獨熱編碼, 因為要將時間(Dates_Hours)還是 datetime 的形式,所以用 datetime 模組將時間 parse 出來才能做 one-hot encoding, 完成特徵工程後,切分訓練跟測試集,建模後將資料丟進去 train,將圖畫出。

再用 y_test 和模型跑出來的結果比較,發現準確度只有

```
In [10]: Acuuracy:  0.22082974584648746
```

學生有兩個推測:

1. 該模型不適合,可改用 RandomForest 等相較於 Decision tree 較有解釋機制的模型。(但還是很爛)

```
In [23]: from sklearn.ensemble import RandomForestClassifier
         %time
         depth_list = [6,7,8,9,10]

         for depth in depth_list:
             rfc = RandomForestClassifier(max_depth=depth, n_estimators=200)
             rfc.fit(x_train, y_train)
             y_pred_rfc = rfc.predict(x_test)
             print(f'Accuracy Score: {metrics.accuracy_score(y_pred_rfc, y_test)}')

         Wall time: 0 ns
         Accuracy Score: 0.21657487255880062
         Accuracy Score: 0.21862942058101342
         Accuracy Score: 0.21918519632094682
         Accuracy Score: 0.22059285782618798
         Accuracy Score: 0.22218729642435756
```

2. 資料集太爛 , 39 種特徵並不適合使用 Decision Tree

圖片參考：附檔 DecisionTree.pdf