團隊測驗報告

報名序號：109005

團隊名稱：C3000

一、資料前處理(說明資料前處理過程)

Step1:確認資料品質

-資料筆數

-遺失值

-資料型態

-資料的特徵數量（欄位數量）

-資料不一致的紀錄方式

Step2:遺失值處理

-將資料內空白值設定為NA(遺失值)

-用回歸模型進行預測 Regression substitution：使用回歸模型來填補缺失值

Step3:文字清理

-將利用文字表達的位置變數欄轉換為距離中心點的距離、角度

Step4:位置變數取代

-將處理完的位置變數取代原本用文字表達的位置變數欄

一、資料前處理(說明資料前處理過程)-續

Step5:資料轉換

-自訂義函數來實現搜尋特定變數，縮小資料集的維度及雜訊

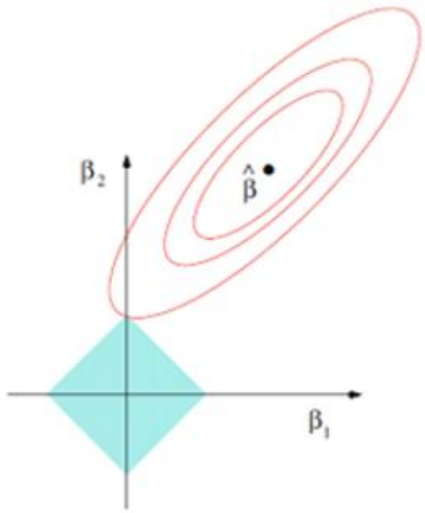
Step6:減少資料量

-透過自訂義的函數來篩選20個重要變數，並透過這20個變數利用Lasso法篩選變數

Lasso Regression:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t. \quad \text{(L1 term)}$$



二、演算法和模型介紹(介紹方法細節)

Lasso Regression

在進行迴歸分析時，我們將要預測的變數當作反應變數，剩下的變數當作解釋變數，透過迴歸分析我們可以去預測反應變數，但這裡會有一個問題，過多的解釋變數，會造成我們SSE提升，造成Overfitting的結果，這種情況是我們非常不願意樂見的，另外一種狀況就是資料本身具有高度共線性，資料之間具有高度的相關，造成迴歸係數的不穩定，這也是我們不想樂見，綜合上述兩點高度共線性以及Overfitting的問題，是高度影響預測結果的原因。

因此要避免上述兩點原因，我們可以藉由Lasso Regression來避免，Lasso Regression為一種同時進行特徵工程和正則化的迴歸分析方法，主要目的為增強統計模型的預測準確性和可解釋性。Lasso Reg-ression是由Regularized Regression所演化爾來，其目標函數與OLS Regression相同，但多了一個稱為 Penalty parameter的參數為$minimize\{SSE + P\}$，而這個參數分別對應兩種分別為L1 Penalty以及L2 Penalty，而Lasso正好對應為L1 Penalty $minimize\{SSE + \lambda\sum_{j=1}^{p}|\beta_j|\}$，Lasso Regression使用Regular-iztion來優化模型，同時也具有變數篩選的功能，因此可以避免上述兩種原因的產生，故我們在這裡利用之。

三、預測結果 (Training Accuracy, Validation Accuracy...)

Training Accuracy and Validation

對於訓練集中的Training Accuracy，我們在訓練集中的利用Lasso Regression 的交叉驗證，利用MSE最小為篩選變數的原則，附表為交叉驗證的MSE表

	Input_A6_024	Input_A1_020	Input_A3_016	Input_A2_016	Input_A3_017
原本標準差	0.01192	0.73462	0.013985	0.013387	0.012766
RMSE	5.41E-06	0.533968	0.000103	7.85E-05	9.84E-05
SQRT(RMSE)	2.33E-03	0.730731	1.02E-02	0.008862	9.92E-03
	Input_A6_001	Input_A3_018	Input_A6_019	Input_A6_011	Input_A3_015
原本標準差	0.038712	0.012696	0.013163	0.002564	0.030817
RMSE	2.48E-05	8.50E-05	8.68E-05	1.45E-06	0.000786
SQRT(RMSE)	0.004982	9.22E-03	0.009315	1.20E-03	0.028039
	Input_A2_024	Input_A3_013	Input_A2_017	Input_C_013	Input_C_046
原本標準差	0.012573	0.001621	0.013847	0.000567	0.000332
RMSE	5.86E-06	2.56E-06	8.64E-05	3.18E-07	7.67E-08
SQRT(RMSE)	2.42E-03	0.001601	9.30E-03	0.000564	2.77E-04
	Input_C_049	Input_C_057	Input_C_058	Input_C_096	
原本標準差	0.000284	0.008267	0.006032	0.007075	
RMSE	5.07E-08	1.53E-05	2.99E-06	4.72E-05	
SQRT(RMSE)	0.000225	3.92E-03	0.001729	6.87E-03	

四、其他(或自行定義項目)