**Machine Learning Assignment 1 Report**

**107062338邱俊維**

此次Assignment我嘗試使用了Linear regression、Auto regression以及ARIMA來進行預測。

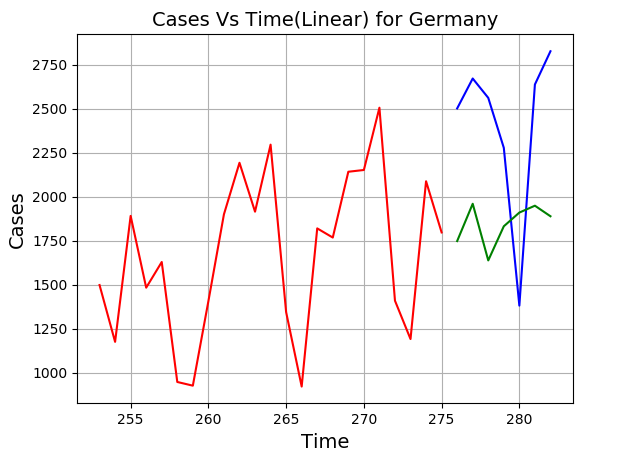
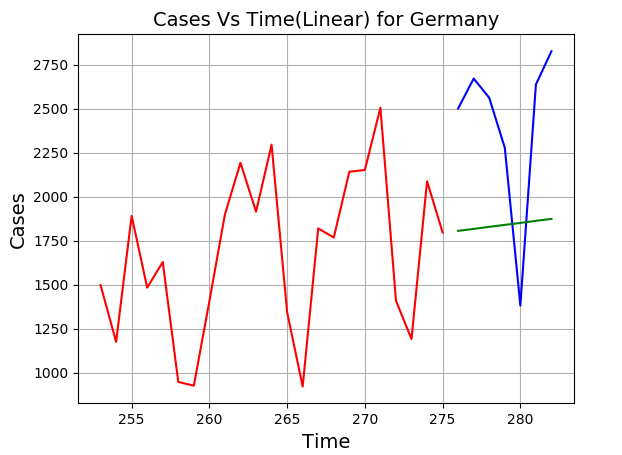
Linear regression、

1. Y = w0 + w1 X
2. Y = w0 + w1\*X + w2\*X^2 + w3\*X^3

+ w4\*X^4 + w5\*X^5 + w6\*X^6 + w7\*X^7 + w8\*X^8 + w9\*X^(1/2)

其效果分別為：

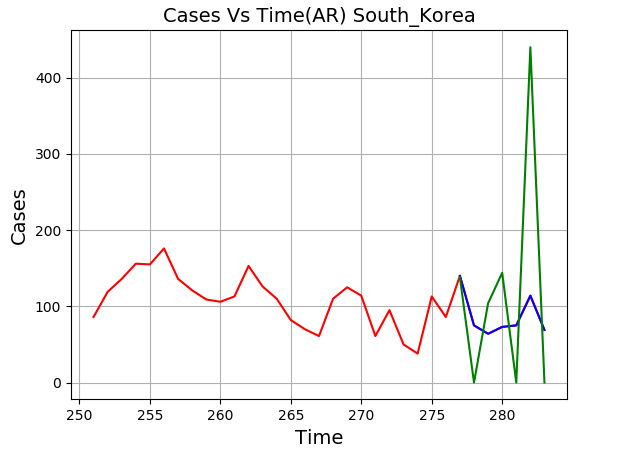
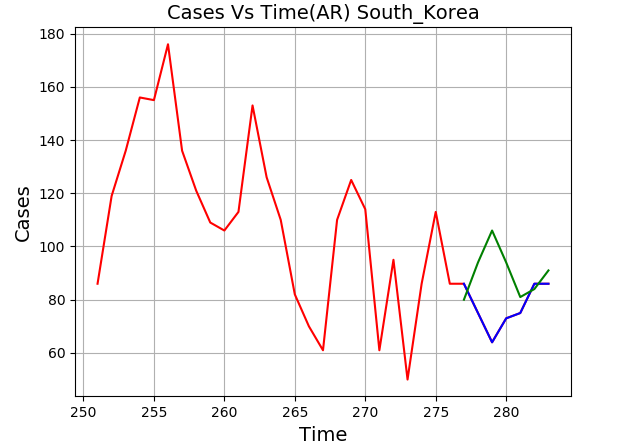
1. (2)



(紅線及藍線為實際cases數，綠線為預測值)

Auto regression、

而在Auto regression方面，本以為這次assignment的資料格式會很適合用Auto regression的方法。但或許是因為data的限制，做出來的效果並沒有比較好，以Linear regression來說，在測試時所有國家的平均MAPE大約為40%上下，而使用Auto regression時卻會上升到45%以上。另外在Auto regression的預測下，偶爾會出現預測值飆升(嚴重者上升到數千萬)的狀況，測試後懷疑可能是data的時間長度抓的不夠長，導致Model因某些Outliers而變形。若拉長時間長度可以避免這樣的情況，但在拉長training的時間長度下，有可能導致MAPE上升。

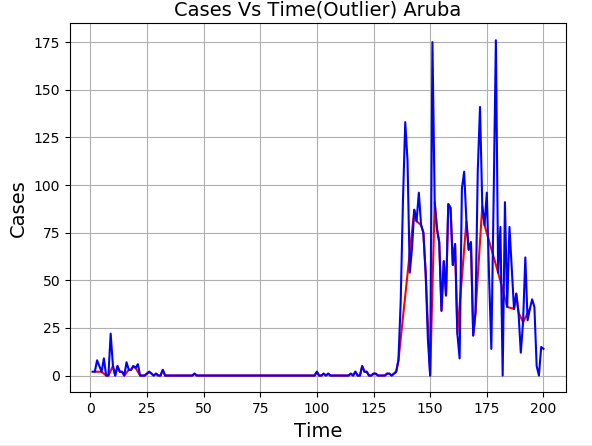
 

左圖取30天作為資料長度 右圖取200天作為資料長度

(紅線及藍線為實際cases數，綠線為預測值)

為了解決在Auto regression裡頭產生的預測值飆升問題，我還去找了過濾掉Outliers的方法，在Python的涵式庫裡找到了IsolationForest來做篩選。



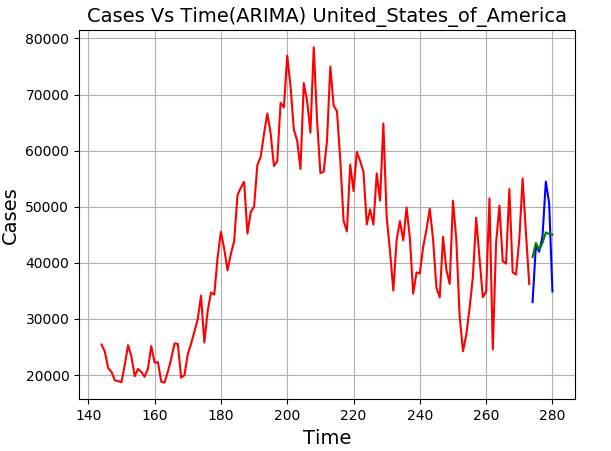


紅線為過濾掉Outliers後的結果，藍線為未過濾。並且可自行設定要過濾掉的比例(下圖設定contamination為0.2)。但在進行過濾後，training的效果依然不好。



ARIMA、

最後還嘗試了ARIMA的方式來做training，但因為不太清楚ARIMA的參數設定方式，於是找到了auto\_arima來自動挑選出最佳的參數。但此方法相當耗時，要產出全部國家的預測結果需要數個小時，也因此到目前為止並沒有做相當多的分析，但在產出的3、4次預測中，平均的MAPE皆略小於40，或許是個值得嘗試的方法。



以ARIMA進行training，預測美國cases數

(紅線及藍線為實際cases數，綠線為預測值)

心得、

很早便著手這次的作業，其實在一開始使用線性(直線)的做法所得到的結果，與在嘗試各種不同方法以及優化後並沒有相差太多。在花費大量的時間找尋更佳的方法後，卻得不到更好的結果，相當的令人沮喪。

因為是初學者的關係，感覺這次作業在資料分析以及優化的部分做的比較不夠，大部分的時間都花費在盲目地調整參數。雖然資料本來就會有高低起伏，產生誤差是一定的，但若能把資料做更好的預處理，我相信還能夠有更好的結果。