LearnPlatform COVID 19 Impact on Digital Learning

R09945067 王馨、R09942113 張凱傑、B07703078 艾芯、R09942172 莊鎧爾

Introduction

一、資料集

我們使用 Kaggle: LearnPlatform COVID-19 Impact on Digital Learning 上的資料及進行分析，資料集中有三子資料集：

1. Engagement data

資料中涵蓋 2020-01-01 至 2020-12-31 的學生使用線上學習平台的使用情況，共有 233 個檔案，每個對應一個地區，每個檔案中每一筆資料涵蓋以下四個了欄位。

time	資料統計日期
lp_id	使用的線上平台 id
pct_access	一天中至少瀏覽一頁特定網站的人數比例
engagement_index	一天中每一千人瀏覽特定網站

2. District information data

地區資料，每個地區有以下的欄位。

district_id	233 個地區對應代號
state	該地區所屬的美國州名
locale	地區類型（City/Suburb/Town/Rural）
pct_black/hispanic	黑人或西班牙裔之學生比例
pct_free/reduced	獲免費或減價午餐之學生比例
county_connections_ratio	網速 > 200kbps 的家庭比例
pp_total_raw	政府對學區補助之中位數

3. Product information data

線上學習平台的資訊，每種線上平台（產品）有以下欄位。

lp_id	產品代號
URL	產品網址
Product Name	產品名稱
Provider/Company Name	發行公司
Sector(s)	使用該產品之教育部門
Primary Essential Function	產品功能分類

二、資料分析方法

在這個專案中，我們分成三個部分對資料進行分析，前兩者我們使用不同方法針對地區資訊以及產品資訊分析其與學生使用線上平台的關係，在分析的過程中我們發現有大量的數據缺失，因此第三部分，我們希望使用機器學習的方式，看看能不能使用前幾天的學生參與指標，預測我們未知的參與指標。

1. 地區特性對於學生參與線上學習的影響

這個實驗中，我們不考慮不同線上學習平台的差異，單純針對不同地區的地區特性對該地區學生對線上學型參與度的影響，其中有地區的所屬洲區、地區類型、人種、午餐補助、網路、政府補助。

2. 線上學習平台預測

這個實驗中我們使用地區以及部分線上學習平台資訊去預測線上學習平台的功能種類。我們把資料以 4:1 的方式分成 training 和 testing data，使用 DecisionTree、Bagging、Random Forests 三種方式預測，分析其預測的精準度可以達到多高，藉此得知學生使用該平台學習時，我們可不可以使用該學生對應的相關資訊預測到該平台的使用功能。

3. 預測各地區的線上學習參與程度

這個部分，我們會使用 regression、one-layer perceptron (OLP)、multilayer perceptron (MLP) 三種模型，嘗試用前幾天的學生參與指標，預測當天的學生參與指標，我們分別使用前 1 日、3 日、7 日、10 日和 14 日的資料去訓練，分析其預測的準確度。

Districts Data Statistics

一、目的：分析不同地區學生參與程度之間的差異，找出造成差異相關性較高的因素。

二、方法：

本實驗中，針對每日線上學習參與程度指標 (engagement_index) 的總和作出統計分析，比較時間與參與程度的關係，並且根據地區資訊，分析這些資料與參與程度的相關性。

1. 資料前處理

在此實驗中，我們不考慮不同線上平台的差異，因此我們把同一地區當日所有參與程度加總，當成當地學生於當日的參與程度指標，因此我們的 engagement data 會轉變為 233 的時間序列，對應 233 個地區 2020-1-1 至 2020-12-31 的學生參與指標，而這 233 個地區有其地區對應數據。

經過資料處理之後，會發現許多地區有大量的資料缺失，造成有些日期的話生參與程度為 0，我們認為這總數據是不合理的，因此移除超過 20 日參與指標為 0 的地區，經過移除之後，會剩下 177 個地區。

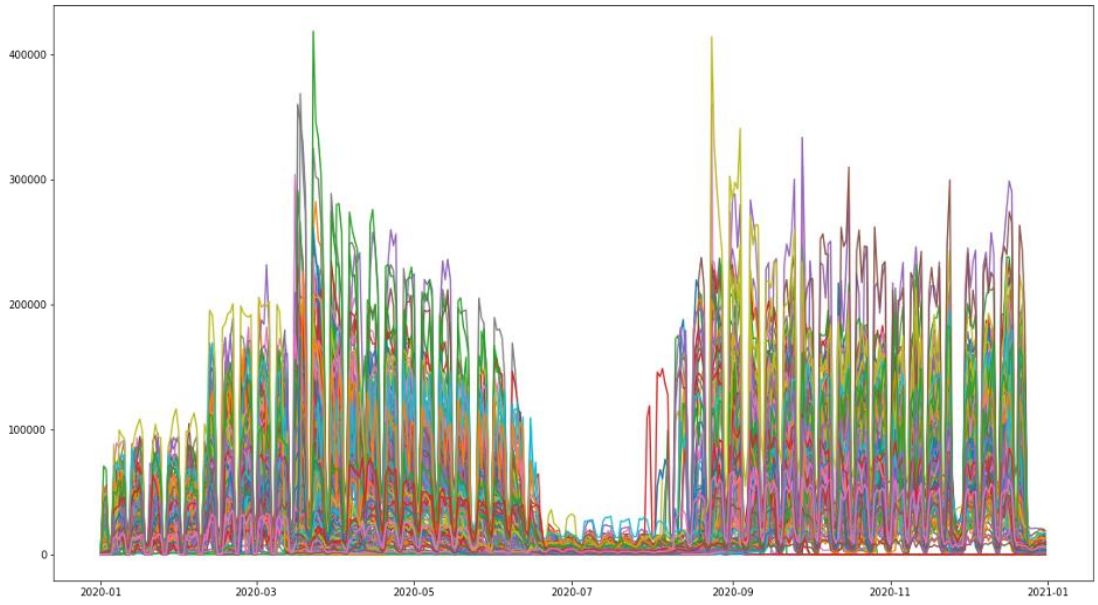
2. 資料分析

我們跟據不同地區數據，將同種類的當成一類別，取平均以及標準差，比較不同種類之間的差異，而人種、午餐補助、網路、政府補助，我們會取相關係數，比較該項地區資訊與學生參與度的相關性有多高，其中的網路資料除了 nan 都是同一個值，因此直接忽略。

三、實驗結果

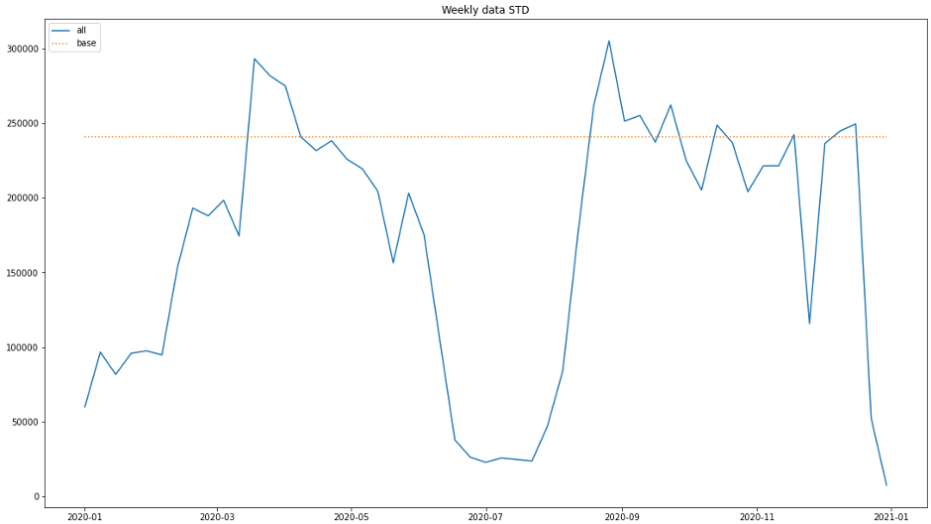
1. 時間與學生參與程度的關係

我們將 177 個地區學生參與指標分別對時間做圖，可以發現學生參與程度與時間有大幅度的相關性，推測是根據學校授課時間造成的差異，導致假日以及暑假時學生使用線上平台學習的參與度大幅度下降，為了減少曲線週期性的震動，之後的數據會以週為單位呈現，避免假日時間出現大幅度參與指標下降的情況發生。



圖：177 地區學生參與指標-時間圖（以天為單位）

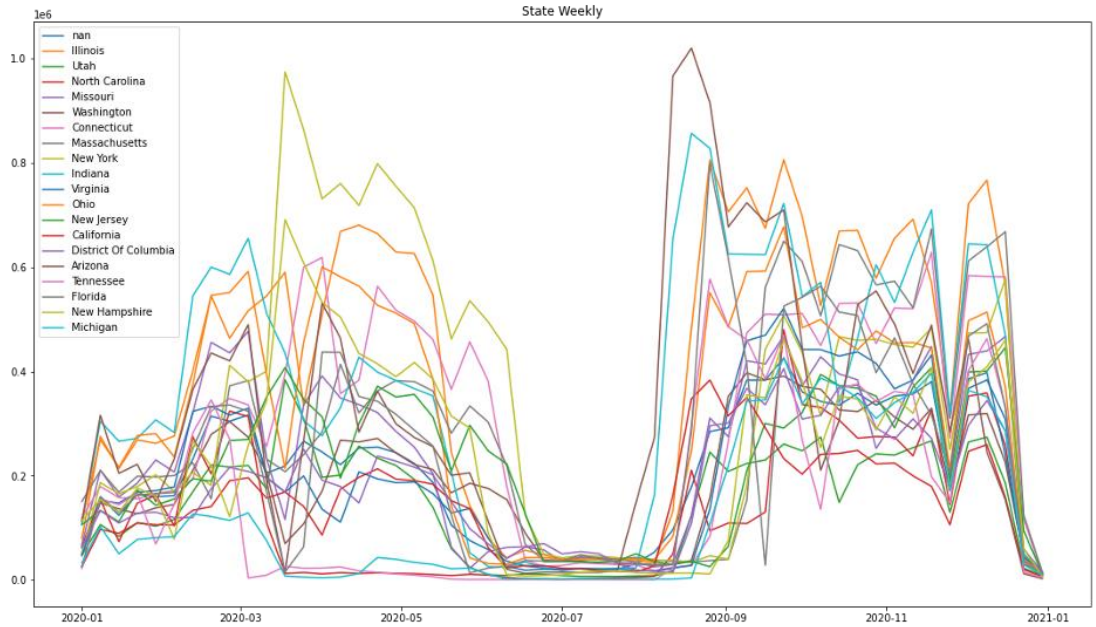
為了體現不同地區在相同時間有較接近的分佈情況，我們對每週的參與指標做不同區域間的標準差，如下圖，虛線是不考慮時間因素，所有數據的區域標準差，實線是當週數據的地區標準差，可以看出，大部分的時間標準差都低於總體標準差，表示不同地區學生參與指標都與時間有相似的分佈。



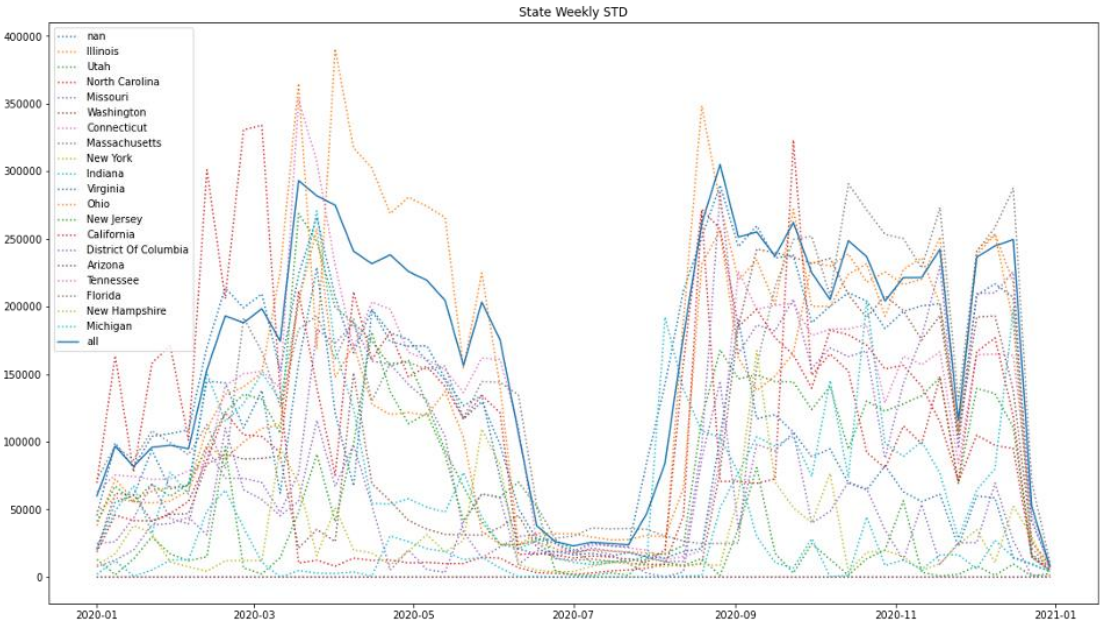
圖：學生參與程度的地區間標準差（以週為單位）

2. 相同州區的學生參與度

177 的地區所屬 19 個州區，我們將相同州區的地區參與指標取平均以及標準差，我把把州區內部的標準差與所有州區的標準差作比較，可以看出，大部分的州區區域內標準差都是低於總體標準差的，因此可以初步判斷，相同州區內的學生參與程度有較接近的分佈。



圖：州區內部的學生參與指標平均（以週為單位）



圖：州區內部的學生參與指標標準差（以週為單位）

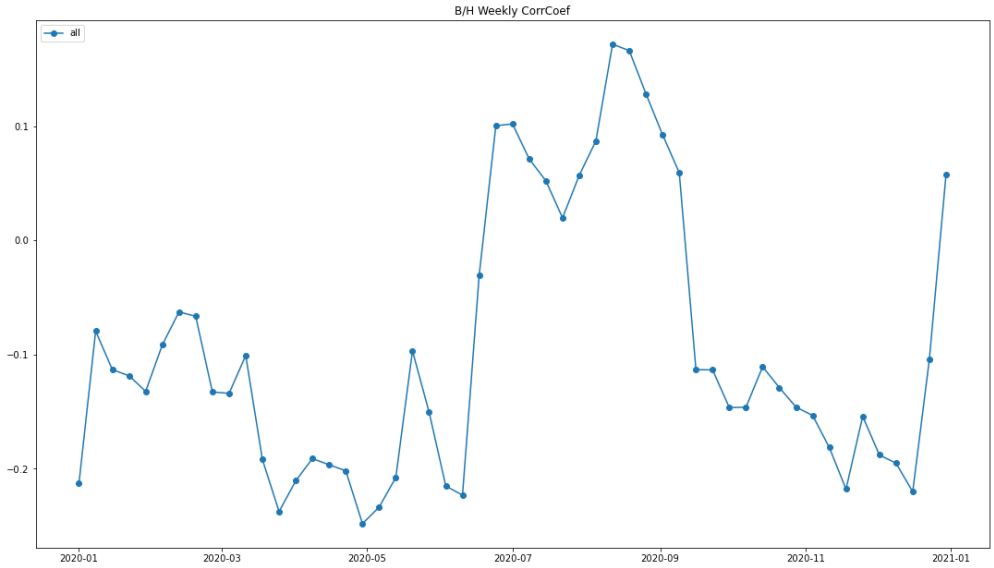
3. 地區種類與學生參與程度關係

地區種類分成 suburb、rural、city、town，我們將相同地區種類的參與指標取平均，可以發現學生參與程度 Rural > suburb > town > city，這與美國的貧富居住區域有關係。

圖：同種地區種類的學生參與指標平均（以週為單位）

4. 拉丁/黑人比例與學生參與程度

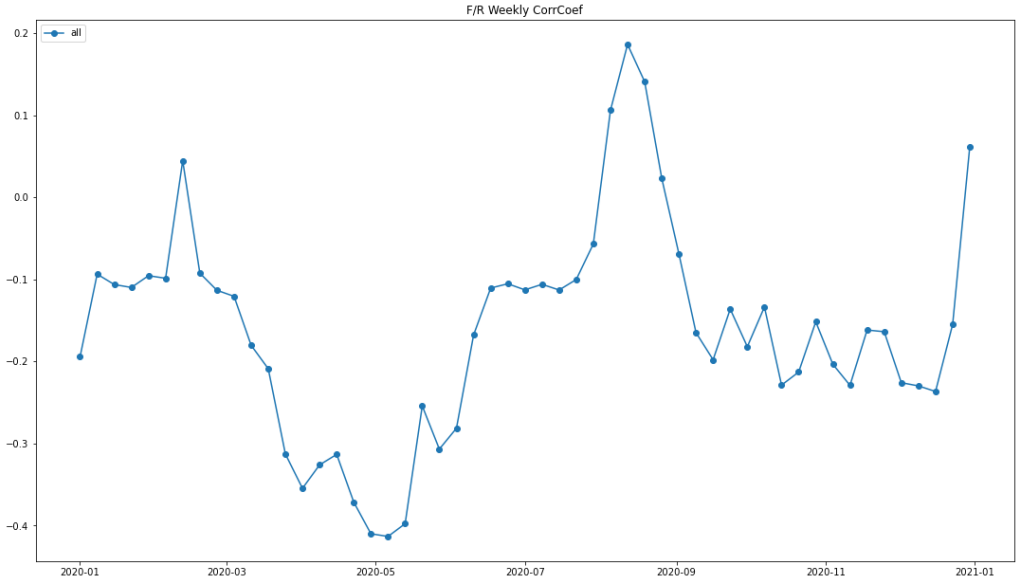
我們把地區的拉丁/黑人比例與學生參與指標做相關係數，如果不考慮時間，相關係數是 -0.0625，幾乎為 0 相關，而對不同週做相關係數，如下圖，可以看出相關係數幾乎為 0 相關。



圖：地區拉丁/黑人比例與學生參與指標相關係數（以週為單位）

5. 午餐補助與學生參與程度

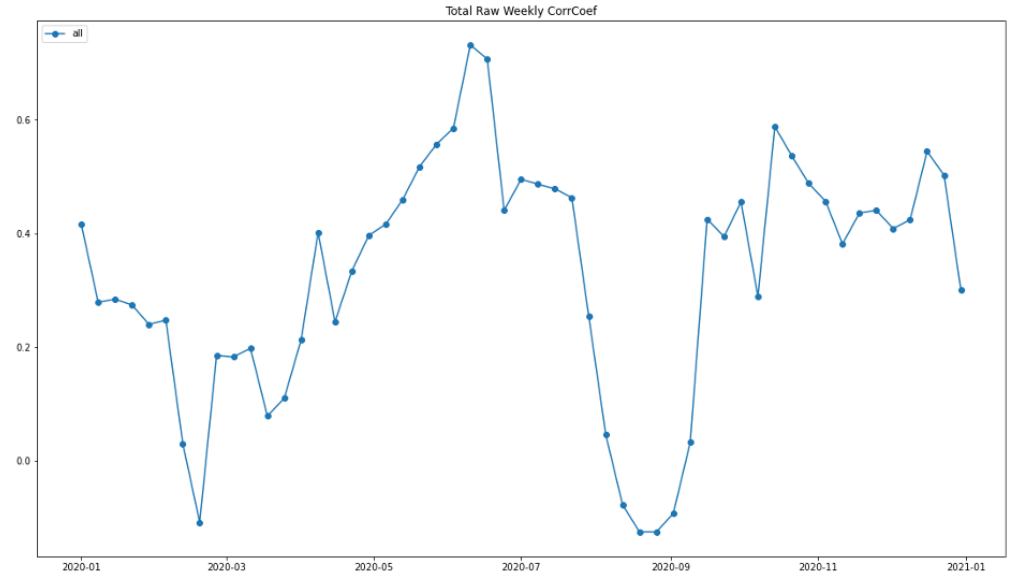
我們把地區的午餐補助比例與學生參與指標做相關係數，如果不考慮時間，相關係數是 - 0.0981，幾乎為零相關，而對不同週做相關係數，如下圖，可以看出相關係數幾乎為零相關。



圖：地區午餐補助比例與學生參與指標相關係數（以週為單位）

6. 政府補助與學生參與程度

我們把地區的政府補助中位數與學生參與指標做相關係數，如果不考慮時間，相關係數是 0.157，為正相關，而對不同週做相關係數，如下圖，可以看出相關係數在學校有授課時是正相關，但是放假時會回到零相關。



圖：地區政府補助中位數與學生參與指標相關係數（以週為單位）

四、結論

1. 數據結果
   a. 學生參與度與休假時間高度相關
   b. 相同州的學生參與度較相近
   c. 學生參與度與居住地有關。

d. 人種、午餐補助與學生參與度無關

e. 政府補助與學生參與度正相關

2. 推論

因為學生大部分都是上課時間使用的線上學習資源，而放假時使用就會大幅度下降，由此推論學生對線上學習資源的使用程度與學校的授課方式有大程度的相關性，因此學生的線上參與度取決於學校或班級對於授課方式的選擇。因此學校政策會是影響學生使用的最重要因素，由數據上也可以看出，只有政府補助與學生參與度有較大相關性，推測是學校選擇授課方式只與地區、政府補助有較大相關性，與人種、午餐補助等等個人補助較無關。

District Engagement Prediction

一、目的：預測各地區的線上學習參與程度。

二、方法：

在本實驗中，我們嘗試使用已知資訊，利用機器學習模型預測特定地區當日線上學習參與程度指標 (engagement_index) 的總和。以下我們將分為資料前處理與模型訓練兩個部份來做說明。

1. 資料前處理：

在取得資料後，我們先將資料進行前處理，以適於訓練。除了將資料進行映射及數值化外，我們也根據所觀察到的特性，加入假期/週末與過往日期的數據作為指標，並在最後進行正規化，以下將逐一說明。

(1) 資料映射及數值化 (Data mapping & numeralization)：

首先，我們收集了各地區的資料，並將其進行資料映射及數值化，如下表。

表：地區資料映射及數值化

Feature Name	**Mapping Method**
locale	City: 4, Suburb: 3, Town: 2, Rural: 1
pct_black/hispanic	[0, 0.2[: 1, [0.2, 0.4[: 2, [0.4, 0.6[: 3, [0.6, 0.8[: 4, [0.8, 1[: 5
pct_free/reduced	[0, 0.2[: 1, [0.2, 0.4[: 2, [0.4, 0.6[: 3, [0.6, 0.8[: 4, [0.8, 1[: 5
pp_total_raw	[4000, 6000[:5, [6000, 8000[: 7, [8000, 10000[: 9, [10000, 12000[: 11, [12000, 14000[: 13, [14000, 16000[: 15, [16000, 18000[: 17, [18000, 20000[: 19, [20000, 22000[: 21,

	[22000, 24000[: 23, [32000, 34000[: 33

(2) 加入假期與週末指標 (Add holiday/weekend features)：

在實驗過程中，由於我們觀察到線上學習參與程度指標 (engagement_index) 與假期和週末之間的相關性相當高，因此我們將假期與週末也加入訓練作為指標。對於美國學生假期，我們參考了 MyKidsTime (https://www.mykidstime.com/school/here-are-the-school-holidays-2019) 和 Edarabia (https://www.edarabia.com/school-holidays-united-states/) 兩大美國教育網站所整理出的美國學生假期資訊。關於我們所參考使用的假期名稱及其對應日期，可參考下表。

表：參考使用的假期名稱及其對應日期

Holiday	Date Start	Date End
Winter break	2019/12/20	2020/1/1
Midwinter break	2020/2/17	2020/2/21
Spring break	2020/4/20	2020/4/24
Summer break	2020/6/22	2020/9/8
Rosh Hashanah break	2020/9/18	2020/9/18
Thanksgiving break	2020/11/26	2020/11/28
Christmas break	2020/12/21	2021/1/1

在假期方面，我們將各假期都視為一個指標，並將假期期間的指標數值設為 1，將非假期期間設為 0。在週末方面，我們將所有週末時間設為一個指標 (holiday)，並將週末期間設為 1，將平日 (即非週末) 期間設為 0。

(3) 加入過往日期的數據作為指標 (Add data from the previous days as features)：

在實驗過程中，我們觀察到線上學習參與程度指標 (engagement_index) 與時間的相關性相當高，因此我們將該地區過往的線上學習參與程度資料也當作指標加入訓練中。對於資料所取時段 (day range)，我們分別取 1 日、3 日、7 日、10 日和 14 日等五種時段來進行訓練。舉例來說，若資料所取時段為 14 日，即

為使用該地區前 13 日的線上學習參與程度指標，加上第 14 日的其餘指標，來預測第 14 日的線上學習參與程度指標。因此，若資料所取時段為 1 日，即沒有使用任何過往日期的數據，僅使用其餘指標 (如地區指標、假期/週末指標等) 進行線上學習參與程度指標的預測。

(4) 資料正規化 (Data normalization)：

最後，我們將資料進行正規化，即將原始資料的數據按比例縮放於 [0, 1] 區間中，且不改變其原本分佈。其公式如下：

$$X_{nom} = \frac{X - X_{min}}{X_{max} - X_{min}} \in [0, 1]$$

其中，X_{max} 與 X_{min} 分別為資料的最小值與最大值，由此即可完成資料正規化。

2. 模型訓練：

做完資料前處理後，我們嘗試使用所整理出的各項指標，來訓練模型預測各地區當日的線上學習參與程度指標 (engagement_index)。我們將資料集分為訓練資料 (training data) 和測試資料 (testing data)，兩者之間的比例為 0.8/0.2。以下，將分別介紹模型基本架構與訓練結果計算。

(1) 模型基本架構：

本實驗使用了三種不同的模型架構來訓練，分別為 regression、one-layer perceptron (OLP) 和 multilayer perceptron (MLP)。其架構分別如下：

```
self.net = nn.Sequential(
    nn.Linear(input_dim,
)
```

regression

```
self.net = nn.Sequential(
    nn.Linear(input_dim, 6
    nn.ReLU(),
    nn.Linear(64, 1),
)
```

one-layer perceptron (OLP)

```
self.net = nn.Sequential(
    nn.Linear(input_dim, 6
    nn.ReLU(),
    nn.Linear(64, 256),
    nn.ReLU(),
    nn.Linear(256, 64),
    nn.ReLU(),
    nn.Linear(64, 1),
)
```

multilayer perceptron (MLP)

在訓練過程中，我們所使用的優化器 (optimizer) 為 Adam，其餘訓練模型參數可見於下表。

表：訓練模型參數表

Parameters	Value
n_epochs	5000
batch_size	128
learning_rate	0.001
weight_decay	0.0005
betas	(0.9, 0.999)
early_stop	350

(2) 訓練結果計算：

在本實驗中，我們以均方根誤差 (root mean squard error，簡稱 RMSE) 作為預測誤差的評價函數，來計算預測數值與實際數值之間的誤差。其公式如下：

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{f}(\boldsymbol{x}^n) - \hat{\boldsymbol{y}}^n)^2}$$

其中 f 函數為訓練好的模型，x 為模型預測出的結果，y 為實際數據，以此計算誤差。RMSE 值越小，則代表模型訓練成果越準確。在本實驗中，我們使用 RMSE 來計算訓練過程中的成果檢驗 (validation) 和最終的訓練成果 (prediction)。

三、結果：

在本實驗中，我們嘗試運用機器學習模型預測各地區當日的線上學習參與程度指標 (engagement_index)，並使用 RMSE 作為誤差的評價函數，以下將根據實驗結果和預測數值兩部分呈現。

1. 實驗結果

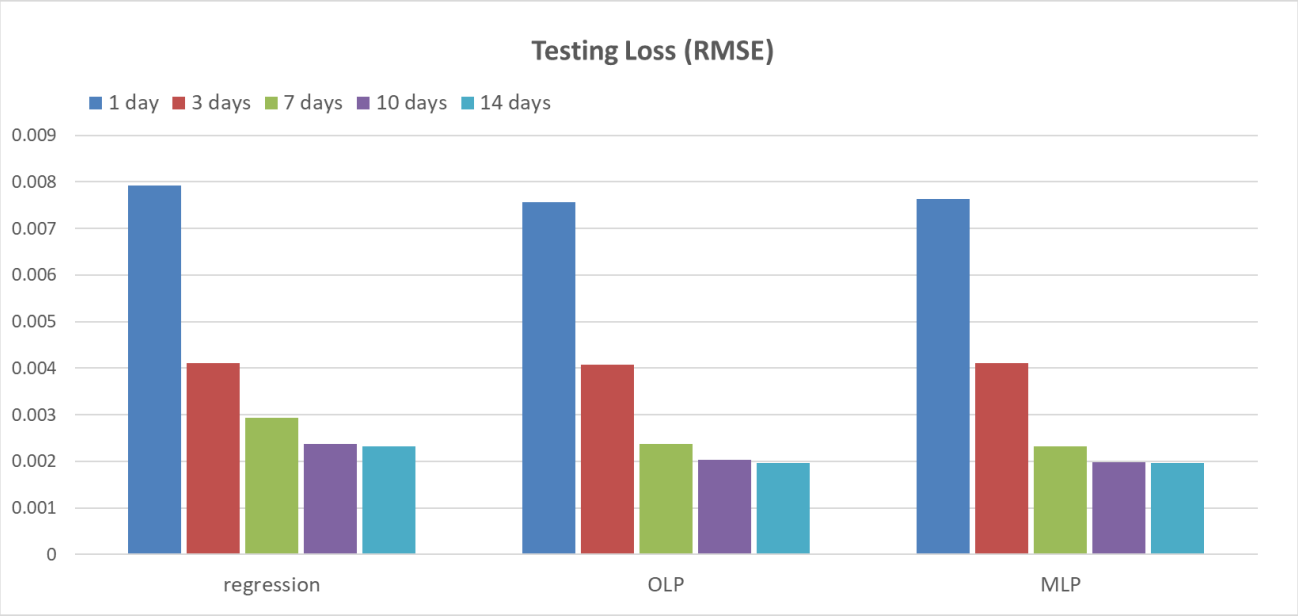
本實驗中，我們所採取的過往資料時段 (day range) 分別為 1 日、3 日、7 日、10 日和 14 日等五種時段。在模型方面，我們分別使用 regression、one-layer perceptron (OLP) 和 multilayer perceptron (MLP) 三種不同的架構來進行訓練，其訓練誤差結果如下表。

表：不同資料時段與模型架構訓練誤差數值

Regression		One-layer perceptron (OLP)		Multilayer perceptron (MLP)	
day range	**testing loss**	**day range**	**testing loss**	**day range**	**testing loss**
1 day	0.007927836	1 day	0.00756923	1 day	0.00764103

3 days	0.004111096	3 days	0.00408301	3 days	0.00410415
7 days	0.002941479	7 days	0.00237948	7 days	0.0023251
10 days	0.002371643	10 days	0.00202358	10 days	0.00198236
14 days	0.002319761	14 days	0.00196099	14 days	0.00197156

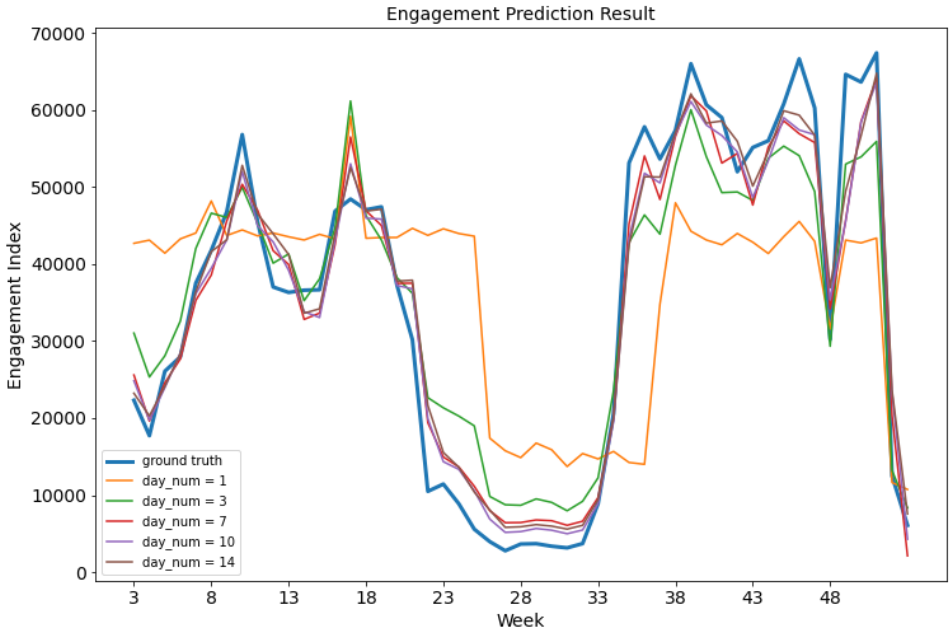
根據上表中的數據，可畫出下圖。圖中，由左至右分別是三種模型架構：regression、one-layer perceptron (OLP) 和 multilayer perceptron (MLP)，而顏色則分別代表了資料時段 (day range) 為 1 天、3 天、7 天、10 天、14 天。由此圖我們可以看出，隨著收集資料的時間越長，預測的準確度就越高。此外，不同模型之間的結果並無太大差異，multilayer perceptron (MLP) 表現最好。



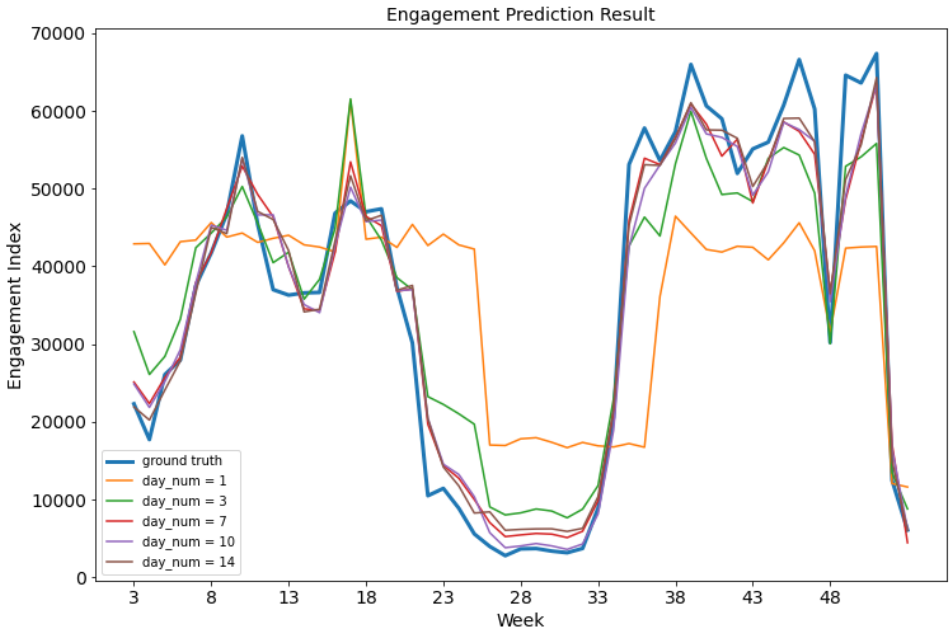
圖：不同資料時段與模型架構訓練誤差圖

2. 預測數值

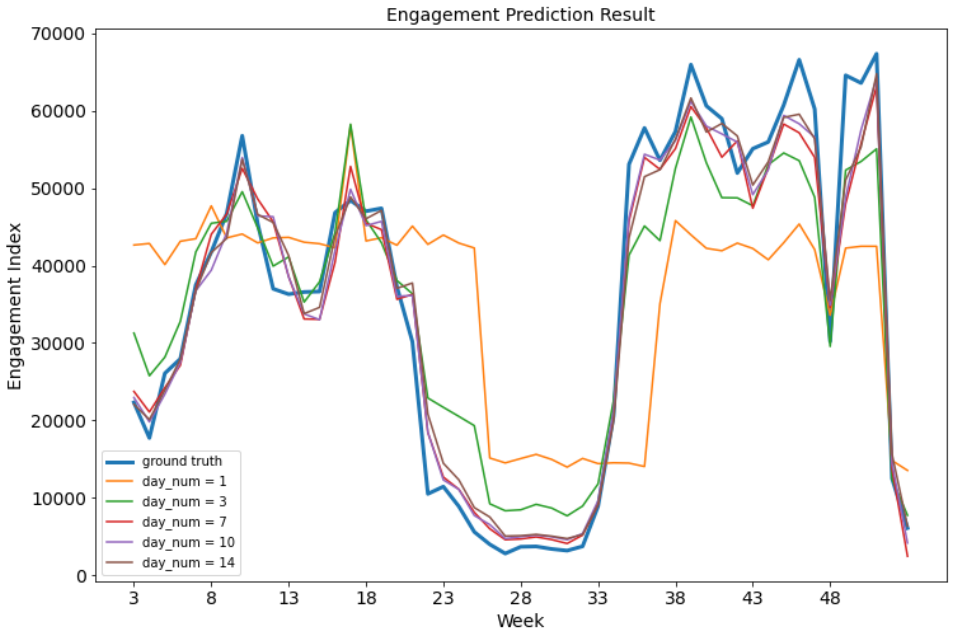
使用所訓練出的模型，我們可以根據各項指標預測出線上學習參與程度。為了更方便檢視結果，使圖形不會過於複雜，我們將資料以七天為單位進行平滑化。結果如下圖至圖，依序分別為三種不同的模型架構。其中，藍線為實際數值，其餘折線則分別代表了資料時段 (day range) 為 1 天、3 天、7 天、10 天、14 天。



圖：實際數值與 regression 模型所預測之結果



圖：實際數值與 one-layer perceptron (OLP) 模型所預測之結果



圖：實際數值與 multilayer perceptron (MLP) 模型所預測之結果

由上圖至圖可看出，資料時段 (day range) 為 1 天的預測結果與實際數值相差較遠，而其餘都與實際數值有高度的相似性。我們推測原因為當資料收集天數為 1 天時，其並無過往數據作為參考，而降低了預測的準確度。由此亦可再次證明，線上學習參與程度指標與時間的相關性相當高。而不同的模型架構之間，並無太大的不同，都可得到相當好的結果。

四、結論

在本實驗中，我們嘗試使用已知資訊，利用機器學習模型預測特定地區當日線上學習參與程度指標 (engagement_index) 的總和。我們使用了三種不同的模型架構，分別為 regression、one-layer perceptron (OLP) 和 multilayer perceptron (MLP)，都可得到相當好的結果。此外，隨著資料收集時段的天數增加，預測結果越好。