Name: Qiyao Zhou
zid: z5379852

COMP9417 Project: LearnPlatform COVID-19 Impact on Digital Learning

1. Introduction

Since its first discovery in Wuhan, China in December 2019, COVID-19 has been affecting the world for close to three years, and people from all countries and regions working in all walks of life have been affected by it, with the education industry being one of them. Most students, including the author, have had to start switching from offline to online learning and, of course, educators have done the same. Many experts and scholars now arguing that COVID-19 has exacerbated the already inequitable state of education.
Less than a year ago, a competition was issued on kaggle with the title LearnPlatform COVID-19 Impact on Digital Learning | Kaggle. This provides data and a platform to analyse the impact of COVID-19 on digital learning, which is the basis for this study.

2. Overview of the problem

This study will explore (1) the state of digital learning in 2020 and (2) how digital learning engagement relates to factors such as district demographics, broadband access, and state/national policies and events based on daily education technology engagement data from more than 200 school districts in the US in 2020.

3. Data set understanding and analysis

3.1 Dataset overview

The overview is prepared to get the feel on data structure. In general, there will 3 datasets: engagement, districts and products where engagement consists of a participation dataset.csv file for each school district.
The engagement_ data folder is based on Learn Platform's Student Chrome Extension. The extension collects page load events of over 10K education technology products in our product library, including websites, apps, web apps, software programs, extensions, eBooks, hardware and services used in educational institutions. The engagement data have been aggregated at school district level, and each file represents data from one school district.
The products_info.csv file includes information about the characteristics of the top 372 products with most users in 2020.
The districts_info.csv file includes information about the characteristics of school districts, including data from NCES and FCC.

3.2 Engagement

The engagement data are aggregated at school district level, and each file represents data from one school district. The 4-digit file name represents district_id which can be used to link to district information in district_info. The lp_id can be used to link to product

information in product_info. The following is a description of each message under this dataset:
time: date in "YYYY-MM-DD"
lp_id: The unique identifier of the product
pct_access: Percentage of students in the district have at least one page-load event of a given product and on a given day
engagement_index: Total page-load events per one thousand students of a given product and on a given day

3.3 Districts

The district file includes information about the characteristics of school districts, including data from NCES (2018-19), FCC (Dec 2018), and Edunomics Lab. In this data set, Learn Platform removed the identifiable information about the school districts. LearnPlatform also used an open source tool ARX (Prasser et al. 2020) to transform several data fields and reduce the risks of re-identification. For data generalization purposes some data points are released with a range where the actual value falls under. Additionally, there are many missing data marked as 'NaN' indicating that the data was suppressed to maximize anonymization of the dataset. The following is a description of each message under this dataset:
district_id: The unique identifier of the school district
state: The state where the district resides in
locale: NCES locale classification that categorizes U.S. territory into four types of areas: City, Suburban, Town, and Rural.
pct_black/Hispanic: Percentage of students in the districts identified as Black or Hispanic based on 2018-19 NCES data
pct_free/reduced: Percentage of students in the districts eligible for free or reduced-price lunch based on 2018-19 NCES data
county_connections_ratio: ratio (residential fixed high-speed connections over 200 kbps in at least one direction/households) based on the county level data from FCC.
pp_total_raw: Per-pupil total expenditure (sum of local and federal expenditure) from Edunomics Lab's National Education Resource Database on Schools (NERD$) project. The expenditure data are school-by-school, and we use the median value to represent the expenditure of a given school district.

3.4 Products

The product file `products_info.csv` includes information about the characteristics of the top 372 products with most users in 2020. The categories listed in this file are part of Learn Platform's product taxonomy. Data were labeled by our team. Some products may not have labels due to being duplicate, lack of accurate URL or other reasons.
LP ID: The unique identifier of the product
URL: Web Link to the specific product
Product Name: Name of the specific product
Provider/Company Name: Name of the product provider
Sector(s): Sector of education where the product is used |

Primary Essential Function: The basic function of the product. There are two layers of labels here. Products are first labeled as one of these three categories: LC = Learning & Curriculum, CM = Classroom Management, and SDO = School & District Operations. Each of these categories have multiple sub-categories with which the products were labeled

4. Data observation and analysis

4.1 Data volume observation

The first thing to do before the data can be analyzed and processed is to know the data size of the three data sets used, with the following results:

```
districts:
Number of rows: 233  Number of columns: 7
products:
Number of rows: 372  Number of columns: 6
engagement:
Number of rows: 5810414  Number of columns: 5
```

The descriptive districts and products dataset is not too complex, but the engagement dataset has a large amount of data.

4.2 Missing value check

In the process of observing the dataset, I found that the frequency of missing values in the dataset is high, which can interfere with the subsequent processing of data analysis, and therefore also has the value of understanding.

```
district_id                 0
state                      57
locale                     57
pct_black/hispanic         57
pct_free/reduced           85
county_connections_ratio   71
pp_total_raw              115
dtype: int64
```

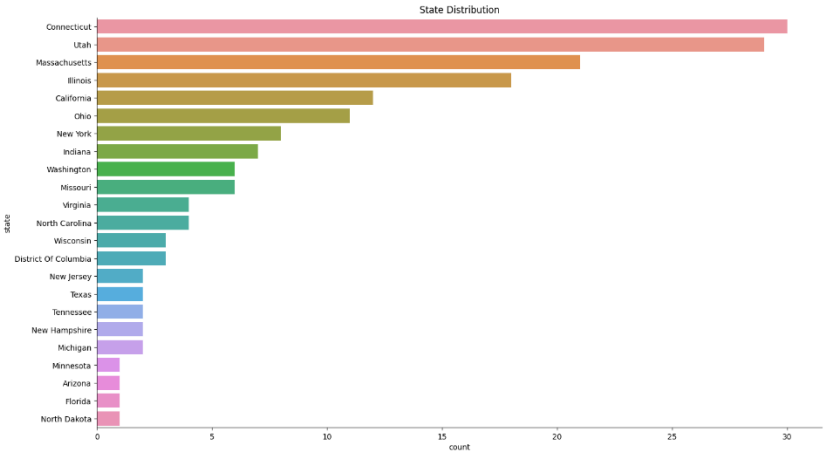
```
LP ID                       0
URL                         0
Product Name                0
Provider/Company Name       1
Sector(s)                  20
Primary Essential Function 20
dtype: int64
```

```
time                        0
lp_id                     114
pct_access               6098
engagement_index      1468404
id                          0
dtype: int64
```

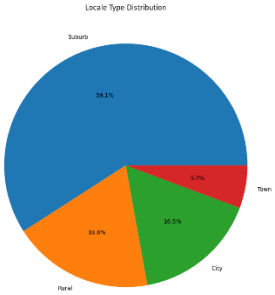
This shows that there are a large number of missing values in the engagement_index dataset, which is the most popular one, and there are some missing values in the district dataset, which need to be taken into account when processing the data.

4.3 Statistical analysis of data

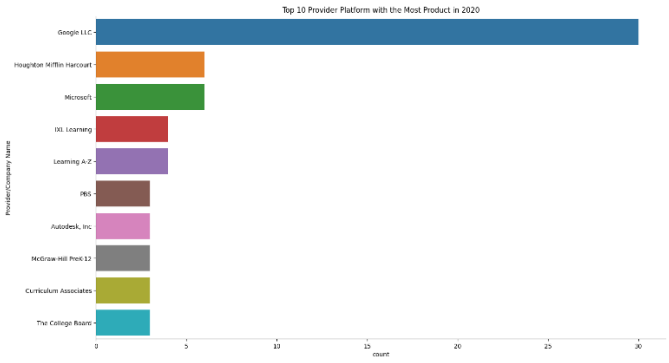
To further understand the characteristics of the data, the next step is to perform a statistical analysis of the data of interest in the dataset to obtain its distribution characteristics.



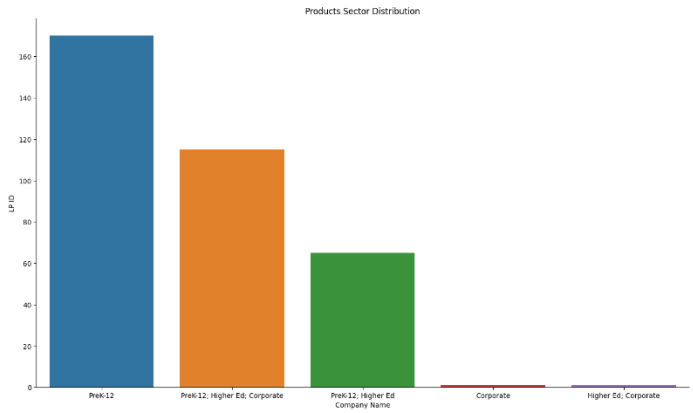
As we can see from the chart above: school districts are not evenly distributed across the states, with Connecticut and Utah accounting for the largest number of districts.



From the data, we know that: Suburb is the highest frequency of locale type while Town is the other way.



Google LLC is a Provider/Company with the most Product and there is no Provider /Company that having products more than 10 other than Google.



From the chart above, we have PreK-12 as the most frequency of platform sector from this dataset. PreK-12 is meant for 1st grade to 12th grade students.

5. Machine learning model attempts

After gaining insight and statistical analysis of the dataset, the next big step was to explore how digital learning participation correlated with regional demographics, broadband access, and state/national policies and events, for which building relevant models for machine learning became the preferred option.

By understanding the dataset, the target value for this study was engagement_index and the features were pct_black/hispanic, pct_free/reduced, county_connections_ratio, pp_total_raw and locale. These data items are distributed in different files so we need to merging the district, product, and engagement data. This can be achieved through district_id and lp_id.

5.1 Attempt at regression model construction

The first model I considered was a linear regression model. Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship between two or more interdependent variables and is very commonly used. In linear regression, there is a linear correlation between the target value and the characteristics, i.e., the equation is assumed to be a multiple primary equation.

Before the model can be formally trained, the data first needs to be pre-processed. Once the desired target and feature columns have been obtained, the first step is to convert the non-numerical features into numerical values. This is done by converting the four columns pct_black/hispanic, pct_free/reduced, county_connections_ratio, pp_total_raw, which are represented by intervals, into the midpoint values of the intervals (e.g. [0.2,0.4] to 0.3), while the locale column is converted into values by urbanisation level by ['Rural', ' Town', 'Suburb', 'City'] to values from 1 to 4.

Next, considering the missing values in the dataset, replace all of them with the mean value of the data in that column.

Then, the data obtained by normalization is divided into a training set and a test set by 0.8:0.2.

Finally, using the squared error as the loss function, the model is optimised using SGD to obtain the desired linear model.

```
[-66.82839788 -18.54295252 -15.10468932   0.          25.62608786]
The train MSE: 16387.418817
The test MSE: 16385.148374
```

From the results of the model training, we can see that the error of the constructed linear model is hardly satisfactory, and since all the data in the county_connections_ratio column in the actual dataset is [0.18,1], which has no effect on the target value, the weights are trained to 0.

So, it is unreasonable to use a linear model to describe how digital learning participation relates to factors such as regional demographics, broadband access, and state/national policies and events.

5.2 Attempt at classification model construction

Since regression models don't work, what happens to classification models? With this in mind, this research will attempt to construct a classification model based on decision tree.

A decision tree is a tree structure (either a binary tree or a non-binary tree). Each non-leaf node represents a test on a feature attribute, each branch represents the output of that feature attribute on a value domain, and each leaf node holds a category. The process of making a decision using a decision tree is to start at the root node, test the corresponding feature attribute of the item to be classified and select the output branch according to its value until it reaches the leaf node, where the category stored in the leaf node is used as the decision result.

Since the values within the target column engagement_index are discrete, they need to be classified before proceeding with the decision tree, and after careful consideration, the values in the target column are divided into 4 categories (very low, low, average, high) according to their distribution in this study.

```
count    25.000000
mean    193.728326
std     150.305589
min      43.112281
25%      84.126890
50%     153.640663
75%     209.533054
max     577.634222
Name: eng_mean, dtype: float64
```

For the data pre-processing part, no normalization of the data is required using the decision tree model, while the missing values will be handled by direct removal, and the rest of the operations are the same as for the linear model pre-processing.

The classification of the resulting model is reported below:

	precision	recall	f1-score	support
average	0.50	1.00	0.67	1
high	0.00	0.00	0.00	1
low	1.00	0.50	0.67	2
very low	0.00	0.00	0.00	1
accuracy			0.40	5
macro avg	0.38	0.38	0.33	5
weighted avg	0.50	0.40	0.40	5

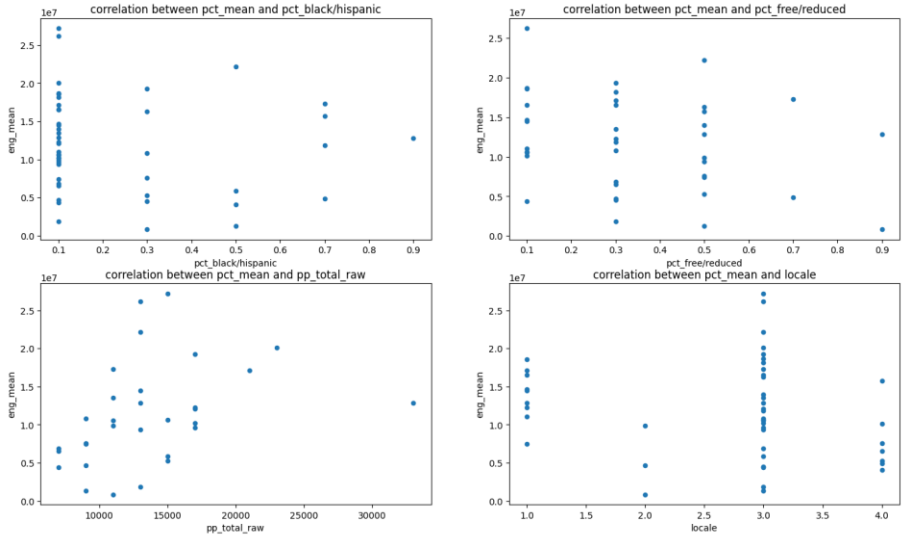
In terms of the results of the resulting decision tree model, the accuracy of the classification results is very low and hardly representative of the general characteristics of the dataset. Therefore, it is also not reasonable to use a classification model to achieve the research requirements.

5.3 Short summary

By experimenting with target regression models and classification models we can see that the machine learning models are not suitable for describing the target-feature relationship required by the study, as evidenced by the low model fit and high error.

6. Result

Based on the results of this study, the relationship between digital learning participation and regional demographics, broadband access, and state/national policies and events cannot be described by simple machine learning models, so examining the simple relationship between digital learning participation and each feature separately becomes the most appropriate solution. The results of the data processing analysis are as follows:



From the results obtained, the above features have no effect on the target values except for pp_total_raw which has some positive correlation with the target. So, enhancing the access to digital device may have impact on increasing engagement rate of student.

7. Conclusions and learnings

In this project, the researcher has basically achieved the task by analyzing the given dataset and attempting to use the basic models of machine learning regression and classification.

The final conclusions show that digital learning participation is only somewhat positively correlated with access to digital devices, but not with regional population, broadband access, and state/national policies and events.

In addition to the results of this project, which are worth considering, this attempt at building a model based on machine learning is also worth summarizing. ML itself can actually be applied by creating a predicting model and checking that if other variables/aggregates have an impact. However, machine learning has certain limits and it is not always a meaningful and necessary tool for data analysis, especially when the target is not clearly correlated with each feature.

8. Reference

Scikit-learn, 2022. Stochastic Gradient Descent. [Online] Available at: https://scikit-learn.org/stable/modules/sgd.html#stochastic-gradient-descent [Accessed 20 July 2022].

Scikit-learn, 2022. Decision Trees. [Online] Available at: https://scikit-learn.org/stable/modules/tree.html#decision-trees [Accessed 20 July 2022].

Nivedita Das K , Aloysius N ,2021,Covid-19: Impact on Education and Beyond Hardcover