2020

WIL PROJECT FINAL REPORT

MACHINE LEARNING BASED COVID-19 CASE ANALYSIS AND PREDICTION

GROUP35 SUPERDS

Yuning Gao S3835212
Yang Liu S3289475
Yang Wang S3646791
Sicong Wu S3735283
Xian Zhang S3749428

Abstract

COVID-19 virus can severe infectious respiratory illnesses. Since December 2019, it has spread to more than 200 countries and has taken more than one million lives in less than a year (World Health Organization, 2020). One of the measures to contain the virus transmission is to lock down the city/ area by shutting down all non-essential places. Despite this strategy's effectiveness in virus containment, the damage to the economy is massive.

This project aimed to explore and identify a measure to control the virus spread while minimising the negative impact on the economy by only shutting down selected places shown to have a higher correlation to the virus transmission. This was achieved by using machine learning techniques

- To develop a machine learning model and predict the future 4th, 5th and 6th days' average daily new cases using the selected date's foot traffic and total confirmed cases as features, based on the assumption that COVID-19's incubation period is 4-6 days.
- To identify places that have a higher correlation to virus transmission and provide advice to decision makers based on the machine learning model's feature importance.

Two datasets of New York State in the U.S., the foot traffic data at various places and the daily COVID-19 cases, were used to develop a prototype. Using the foot traffic and total confirmed cases as features, more than 10 different models were trained and evaluated. The model had the best performance was Adaboost regression. It was based on voting regression that contained the SVM model and KNN model. The Root Mean Square Error (RMSE) of the final model was 234.21, while the Mean Absolute Error (MAE) was 151.84. A web-based interactive tool was developed to provide the users an opportunity to use foot traffic data to predict average new cases in the future; by adjusting the foot traffic data input, users can also see the impact of foot traffic on the virus case prediction in an intuitive way.

The correlation between the foot traffic and the virus transmission was analysed using fine-tuned Random Forest Model. The three places shown by the model to have the highest correlation to new confirmed cases prediction were Auto Shops, Buss, and Fast Food Restaurants.

This project, using the data of New York State, is a prototype that can be used by the government for well-informed decision makings and developing virus control strategies with less economic impact. It has the potential for wider applications with other area's datasets available. More research and analysis need to be completed before applying the model to other areas as the COVID-19 transmission mechanism is complex.

Table of Contents

1 Introduction

1.1 Background

COVID-19 is an infectious illness that is now a global pandemic (World Health Organization , 2020). It is caused by a new form of coronavirus that was firstly reported in December 2019 (Australian Government Department of Health , 2020). This disease can cause severe pneumonia symptoms such as fever, cough, shortness of breath, etc. and can be spread easily from person to person (U.S. Centers for Disease Control and Prevention, 2020). To the date of 10 October 2020, there are approx. 36.3 million accumulated confirmed cases across the world (World Health Organization , 2020). The U.S., in the centre of this pandemic, has more than 6 million confirmed cases in approximately 200 days from 22 Jan 2020 to 5 Sep 2020 as shown in Figure 1.1.

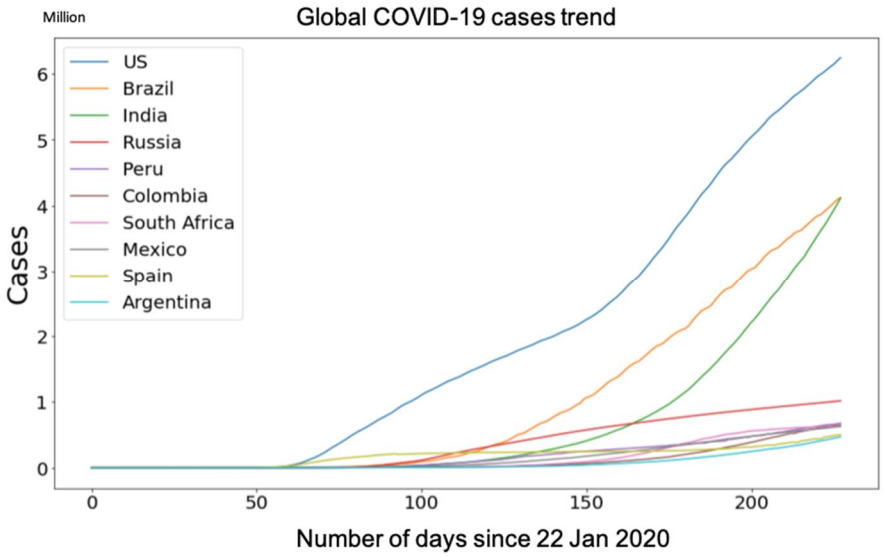


Figure 1.1 Global accumulated COVID-19 cases

Figure 1.2 shows the total confirmed cases in different states in the U.S. over 8 months time.

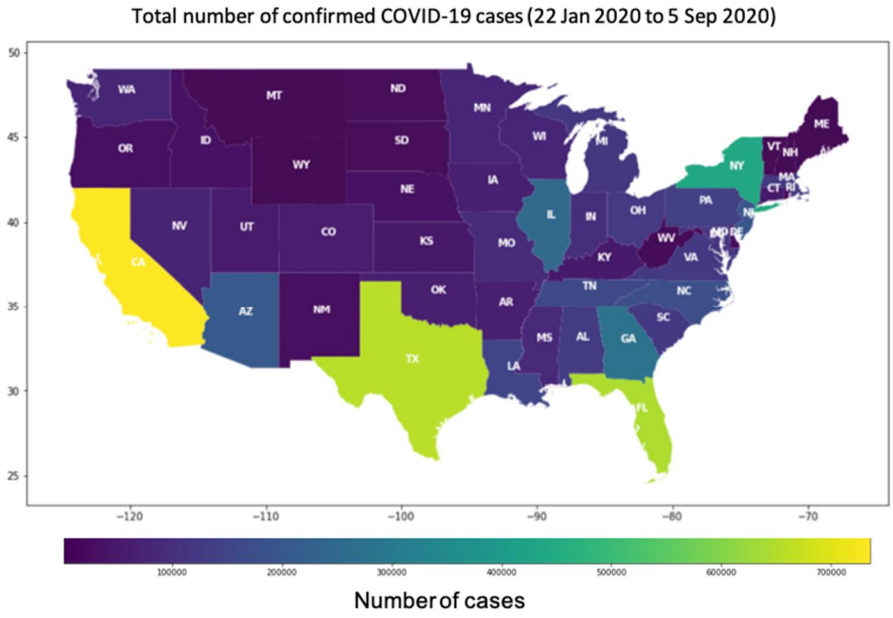


Figure 1.2 Confirmed COVID-19 cases in the U.S. states

1.2 Problem identification

To minimize the transmission and reduce the pressure on health systems, different measures and strategies have been applied across the world. The lockdown control measure has shown its effectiveness in the virus containment (Dickens, et al., 2020). However, the hard lockdown strategy, which requires close-down of all non-essential businesses and restricts people's mobility heavily, is a double-sided sword. The pandemic has already damaged the economy across the world, with IMF predicting the global economy will have a negative 4.9 percent growth in 2020 (International Monetary Fund, 2020). The complete lockdown strategy has more limitations and therefore is expected to hit the economy even worse. In Victoria, Australia, stage 4 lockdown is predicted to result in an economic loss of $9 billion (ABC news, 2020).

While social distancing and lockdown are critical for transmission control (U.S. Centers for Disease Control and Prevention , 2020), closing all non-essential business might not be the optimal solution for the economy. Besides, different places may have different risks in terms of COVID-19 transmission. Some places, e.g. parks, are typically in the open-air and have better ventilation, and some places such as shopping malls are of less effective ventilation and may be riskier. On the other hand, the average time people tend to spend in different places also varies and this can change the risk levels of people getting contracted.

Therefore, it is important to explore the relationship between the foot traffic of different places and the COVID-19 transmission, and develop a strategy that balances the COVID-19 containment and economic damages.

2 Project objectives

The purpose of this project is to explore and identify a virus containment strategy that has reduced negative impact on economic than hard-lockdown by only shutting down selected places that proven to have higher correlation to virus transmission.

The objectives of this project are to use data science and machine learning techniques to

- predict the number of daily new confirmed cases and develop an interactive tool for prediction;
- identify the places that have higher correlation to coronavirus transmission.

3 Proposed solutions

As described in Section 1, social distancing and crowd density have impact on the COVID-19 spread. Research showed the average COVID-19 incubation period is 4-6 days (Lauer, et al., 2020). This means if one gets contracted when interacting with others at certain venues, the test result will likely to be positive 4-6 days later. Therefore, to achieve the project objectives, New York State's data was used to build a prototype. This prototype used foot traffic at different places and total confirmed cases to date as features to predict the average daily new confirmed cases of 4-6 days into the future, and to identify places that have a higher correlation to virus spread.

The following two key datasets were used in this project

- USA main counties' foot traffic data during COVID-19 period provided by Foursquare (Foursquare, 2020);
- USA main cities' daily COVID-19 cases provided by New York Times (Rearc, 2020).

4 Key assumptions

The following assumptions were made in this project:

- The incubation period for COVID-19 is 4-6 days (Lauer, et al., 2020);
- The change of methods of original data collection, e.g., the U.S. COVID-19 case data collection moved from Health and Human Services (HHS) to Centers for Disease Control (CDC) (Medpage today, 2020), did not affect the consistency of the data.

5 Methodology

Machine Learning techniques were used in this project. As shown in the flowchart in Figure 5.1. The project went through stages of project objective identification, data collection, data preprocessing, feature engineering, model training and selection, interactive tool development and conclusions.

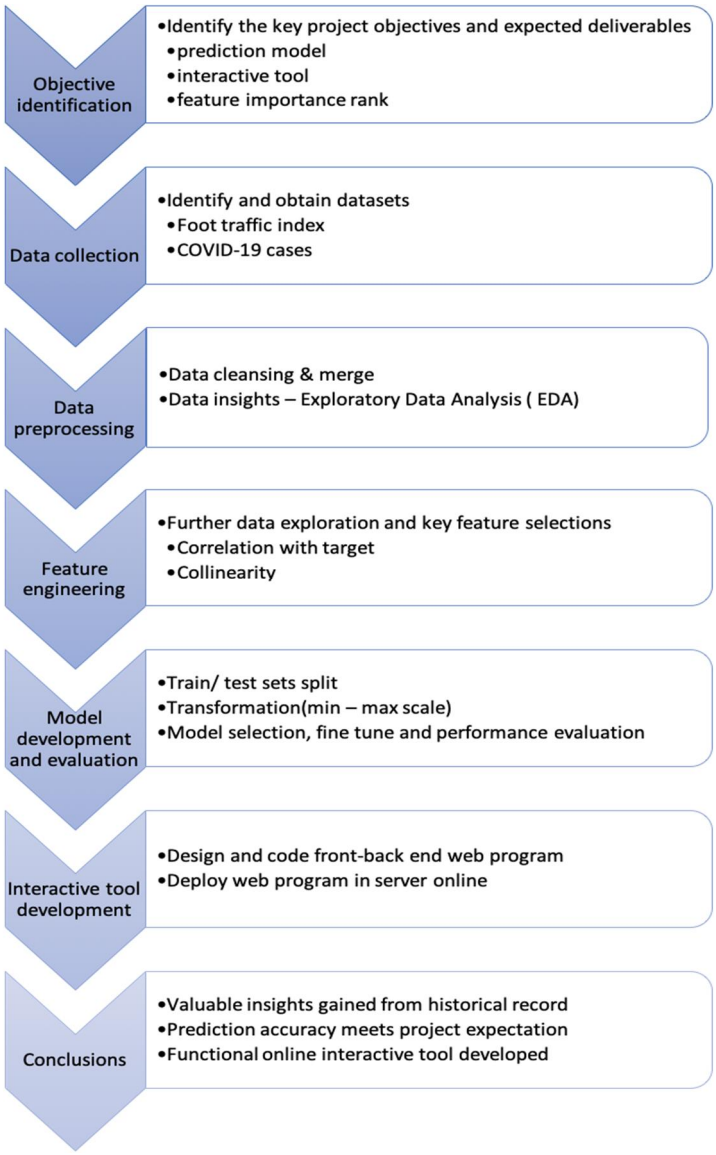


Figure 5.1 Project task flow chart

5.1 Data collection

There are two key datasets used in this project.

- USA main counties' foot traffic data during COVID-19 period provided by Foursquare (Foursquare, 2020) referred to in this project as foot traffic data.
    - This dataset is in csv format.
    - This dataset includes indexed daily foot traffic at the following locations:
        - Airports
        - Auto Dealerships
        - Banks
        - Bars
        - Big Box Stores
        - Casual Dining Chains
        - Clothing Stores
        - Convenience Stores
        - Discounted stores
        - Drugstores
        - Fast Food Chains
        - Gas Stations
        - Grocery Stores
        - Gyms
        - Hardware Stores
        - Hotels
        - Movie Theaters
        - Offices
        - Shopping Malls
        - Warehouse Stores
    - This dataset started on 19 February 2020.
- USA daily COVID-19 cases provided by New York Times (Rearc, 2020), referred to in this project as COVID-19 data.
    - This dataset contains the cumulative counts and new confirmed case counts of coronavirus cases in the United States, at the state and county level over time.
    - The dataset has been updated regularly. In this report the dataset last updated on 28 August was used.

5.2 Data preprocessing and insights

5.2.1 Data preprocessing

The datasets obtained were first preprocessed to resolve issues in the raw datasets such as missing values, unreasonable/ wrong values, etc. The foot traffic datasets contained quite a significant part of missing values for certain venue attributes; these missing values were removed to avoid the impact of insufficient sample size. The New York's relevant data was extracted from each of the datasets, then the two datasets were united as one combined dataset using the mutual key, date. This project focused on the data from March 1, 2020 to August 28, 2020.

Each row contained a date, foot traffic data at 19 venues of the day, total accumulated COVID-19 cases of the day, and the new confirmed cases of the day. In addition, a column of target variables was added. This target variable was the next 4-6 day's average new confirmed cases. For example, for the data of 01/08/2020, the target column's value was the average new confirmed cases of 5/08/2020, 06/08/2020 and 07/08/2020.

5.2.2 Data exploration

In order to have a general understanding of this combined dataset to analyse, the data was visualised using RStudio. This allowed us to effectively observe the foot traffic at different places in New York while observing the trend of new cases average in the next 4-6 days. A snapshot of the visualisation page is shown in Figure 5.2, the link to the visualisation is: https://yang-liu9281.shinyapps.io/projectfinal/, and a sample sheet of the programming code is included in Appendix A.

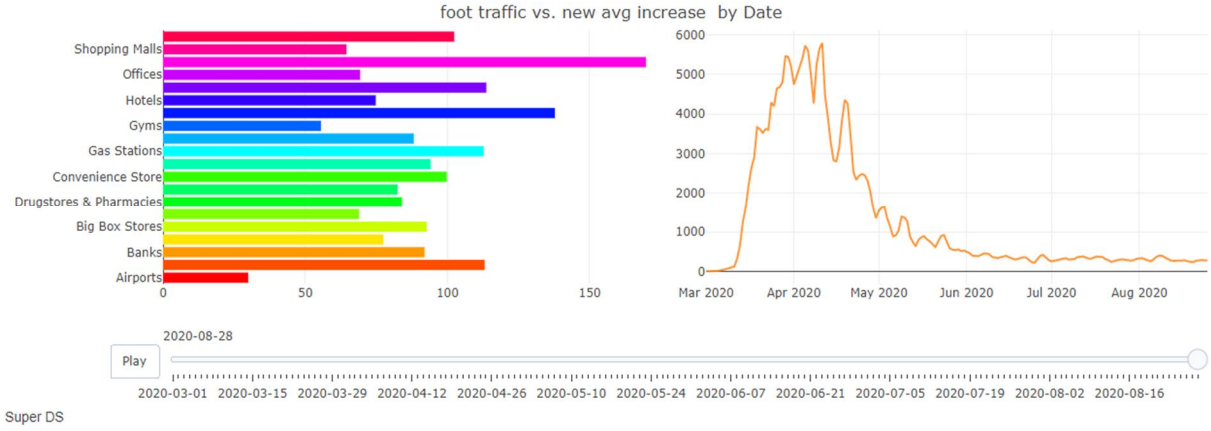


Figure 5.2 Data visualisation demo window

The bar chart on the left hand side shows the daily foot traffic at different venues of a selected date. The line chart on the right hand side shows the average new cases of the next 4-6 days of the selected date. By looking at the trend chart on the right and hovering the cursor over different dates, it can be found that the average number of new cases in New York increased explosively from March 11, 2020 to a peak on April 11, with some minor fluctuations between March 30 and April 11. Then it started to decrease gradually after April 11 with occasional bouncing, then began to decline steadily and plateaued at around 300 new cases daily.

It is clearly shown in the line chart that the average daily new cases have increased at the beginning, then started to decrease with the peak value in mid-April.

Four representative dates were chosen, one at the start of the outbreak, one in the developing period, one at the peak period and one at the low value stable period, to compare and explore the relations between foot traffic and new confirmed cases as shown in Figure 5.3.

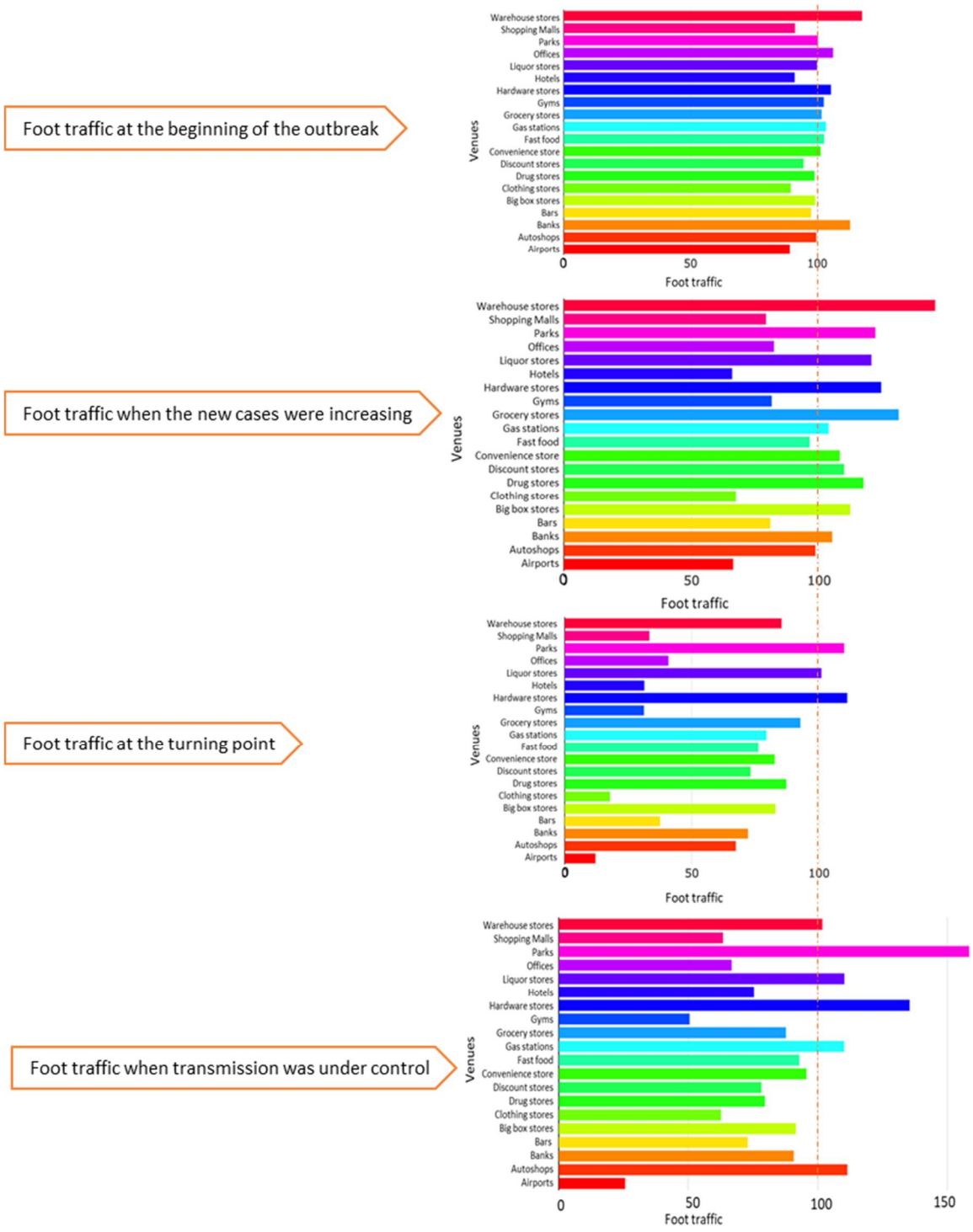


Figure 5.3 Typical foot traffic at different stages of the outbreak

It is shown that at the beginning of this pandemic, the foot traffic at these venues were similar to the before pandemic baseline stats (100); along the time, the foot traffic at certain places such as hotels and airports become lower, probably due to the travel restrictions, while foot traffic at warehouses largely increased. When the foot traffic at many locations reached their lowest record, the newly confirmed cases reached its peak value and the turning point appeared. Then the foot traffic at most of the locations gradually increased but remain lower than the before pandemic value, and the virus transmission was generally under control.

Therefore, it is clear that the foot traffic did affect the virus transmission among other factors such as the accumulated confirmed cases.

5.3 Feature engineering

Considering the findings in Section 5.2, and the further analysis completed on the combined dataset such as visualising the data using scatter matrix , all of the foot traffic data at 21 venues were chosen as features. Also, considering the mechanism of virus transmission, the attribute of accumulated cases was also selected as features for that the larger the size of the contracted group, the higher the chance the virus transmits. After diagnosing the collinearity between each pair of attributes, 22 attributes were selected as features for the prediction model.

The target attribute, as described in Section 5.2.1, the next 4-6 days' average new confirmed cases of the selected day.

5.4 Model selection, training, and fine-tuning

As the project requires the prediction of future cases, model development using Python focused on the regression models following the steps below:

- Selecting individual regression model

  Initially, individual regression models with different algorithms such as the Least Square (with penalties / regularization) model, Support Vector Machine model, K Nearest Neighbours model and Decision Trees model were selected and trialled out. The aim was to identify the combination of hyperparameters for each model and identify the best performed model on the train set cross validation. This best performed model was then used as the base estimator for the ensembled model.

- Selecting ensembled regression model

  Once the best performed individual regression model was identified, ensembled models were developed based on the base estimator/s with different ensemble method, such as bagging, pasting and boosting. Random Forest, voting, Gradient Boost and AdaBoost were considered. In order to have better prediction outcome, voting regression was put into AdaBoost regression. Afterwards, all ensembled models were evaluated via cross validation on train set to came up with one performing best. This model was then evaluated on the test dataset. In addition to developing and fine tuning models for prediction, Random Forest model was used to find the feature importance to provide the factor correlation between foot traffic at different venues and average daily new cases.

- Selecting final model to deploy

  After evaluating on the test data set, the prediction model was identified by comparing RMSE (root mean square error) and MAE (mean absolute error) as lower RMSE and MAE means more accurate prediction.

A sample sheet of the model development code was included in Appendix B.

5.5 Interactive tool development

Once the prediction model was finalised an interactive tool using this model was developed. This interactive tool was programmed based on the agile concept and following the steps below:

Platform selection. Web-based app was selected based on the idea of one-develop, multi-platform versatility. This reduced the development cost and time, and allowed the timely feedback to the rapid iteration of the version. This was ideal for small teams like this project team.

Backend programming. The machine learning model was coded in python, therefore python was also used as this app's back-end programming language to avoid hybrid programming. There are a number of different web frameworks under Python. After comparing many python web frameworks, mature Django was chosen as the web app platform because Django is one of the heavy hitters and has been used for developing many successful websites and apps. Django uses the Model, View and Template (MVT) software design pattern, and

- The Model part is used to interact with the database.
- The View part is mainly used to write the logic of application. We packaged the machine learning model with a pickle, and then called the interface directly from inside the view.py file by the unpackaging pickle file.
- The template part is to store html file for web applications. In this project, Django's own sqlite database was used. This completes the entire back-end build.

Front-end programming. An html page was put together using Vue framework, and ajax of javascript script in which the front and back data interact was added.

For web app deployment, the process of Docker, Django, Gunicorn, Nginx was followed. Firstly, the whole Django project was put into a screen order in a Docker container. Secondly Gunicorn was used to launch this program. Finally, the setting document of Nginx was modified, and Nginx was put to listen to the port where Gunicorn deploys the program.

Once these were completed, the entire program was released as a Minimal Viable Product (MVP). Users can access and use the app directly through the IP or domain. On this basis, the app's functions can be further refined, and the user interface can be improved and djusted to more devices according to the users' needs. Also, in the background, new models were trained based on the latest data and hotfixed without rebooting the project.

6 Results

6.1 Final prediction model

The machine learning model was evaluated and selected as described in Section 5.4 and as illustrated in Figure 6.1.

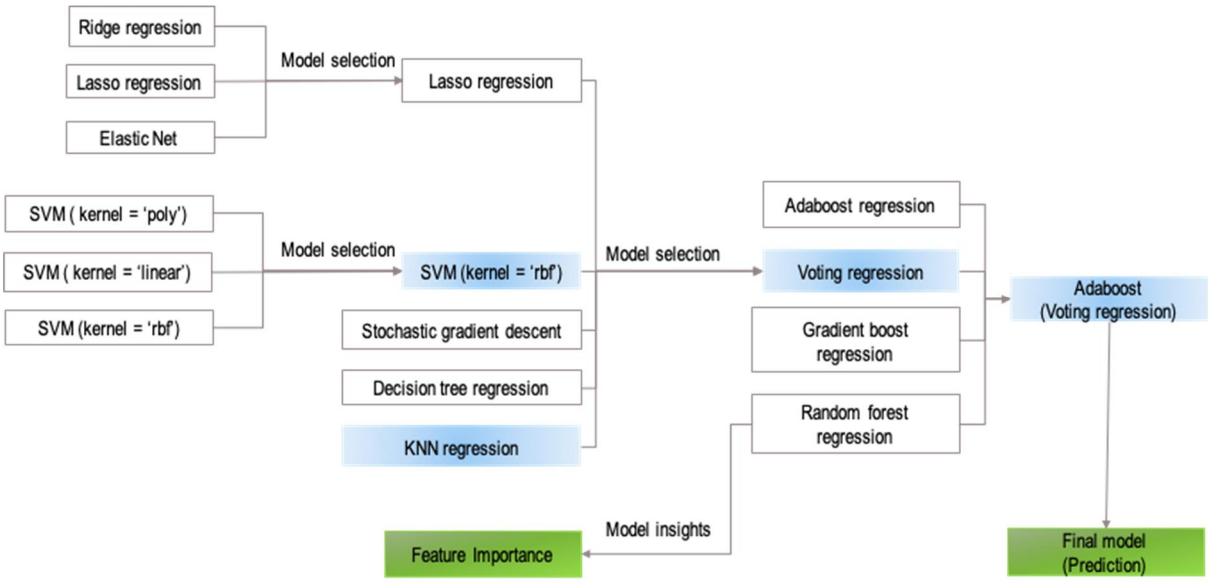


Figure 6.1 Model selection and evaluation

In the step of selecting individual model, SVM with Gaussian Radial Basis Function kernel among Support Vector Machine, and Lasso Regression among regularized linear models were selected. After fine tuning with grid search and random search approach, and comparing the performance of the models using the cross validation, SVM (RMSE: 289.61) and KNN (RMSE: 266.72) turned out to be the top two best performed models. Similarly, for the ensembled models, after comparing the RMSE and MAE. Voting regression (RMSE: 264.5) and AdaBoost regression (RMSE: 258.85) were considered as candidate models for the final evaluation on test dataset. After final evaluation, AdaBoost regression was selected as it had the lower RMSE and MAE on the test dataset compared with voting regression. Figure 6.2 is an example of comparing the predicted new average cases and the actual average cases using AdaBoost on the training data set.

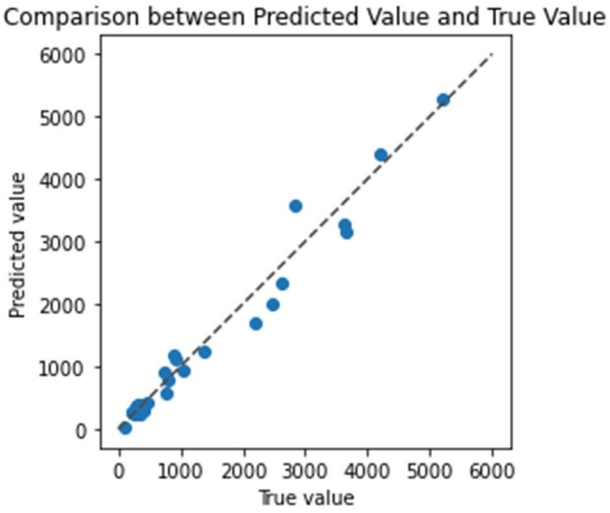


Figure 6.2 Example - final model performance

6.2 Interactive tool

An interactive tool was developed based on the Adaboost model. As shown in Figure 6.3, by entering the foot traffic data at different places, the program will automatically output the average new confirmed cases of the

next 4-6 days based on the current date and the accumulated confirmed cases of the day. By adjusting the input values, users can have an understanding of how foot traffic can impact the new cases in an intuitive way.

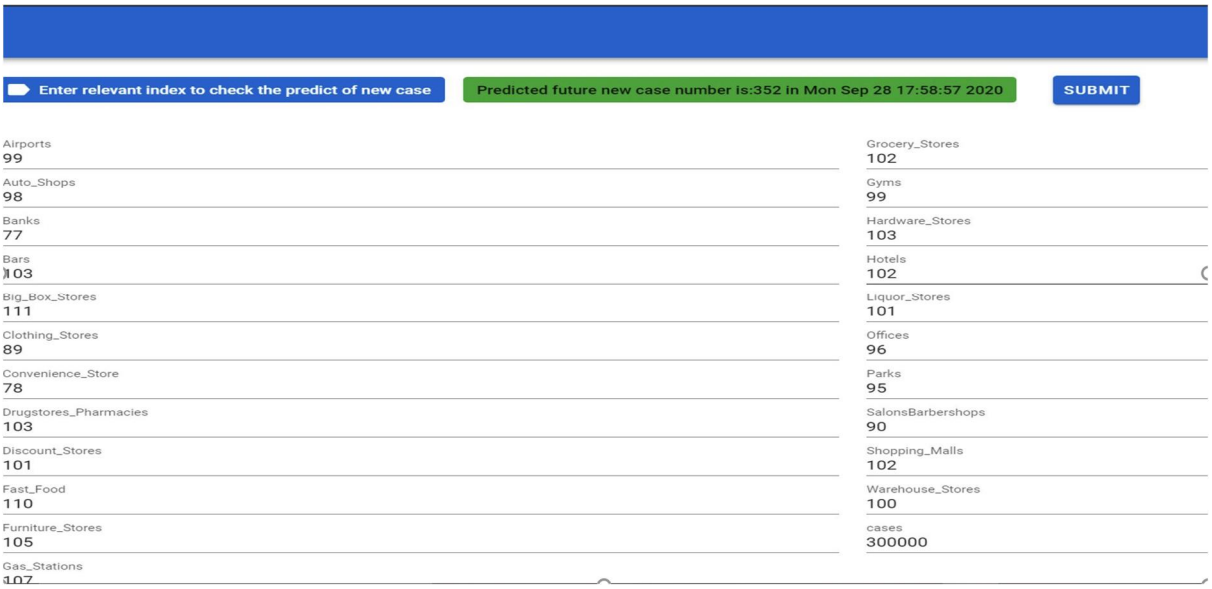


Figure 6.3 Interactive tool screenshot

The link to the interactive tool is: http://155.94.133.17

6.3 Feature's correlation to the new case

The Random Forest model showed that Auto Shops, Fast Food and Bars are the top three locations which contributed most to the new daily COVID-19 cases, as shown in Figure 6.4

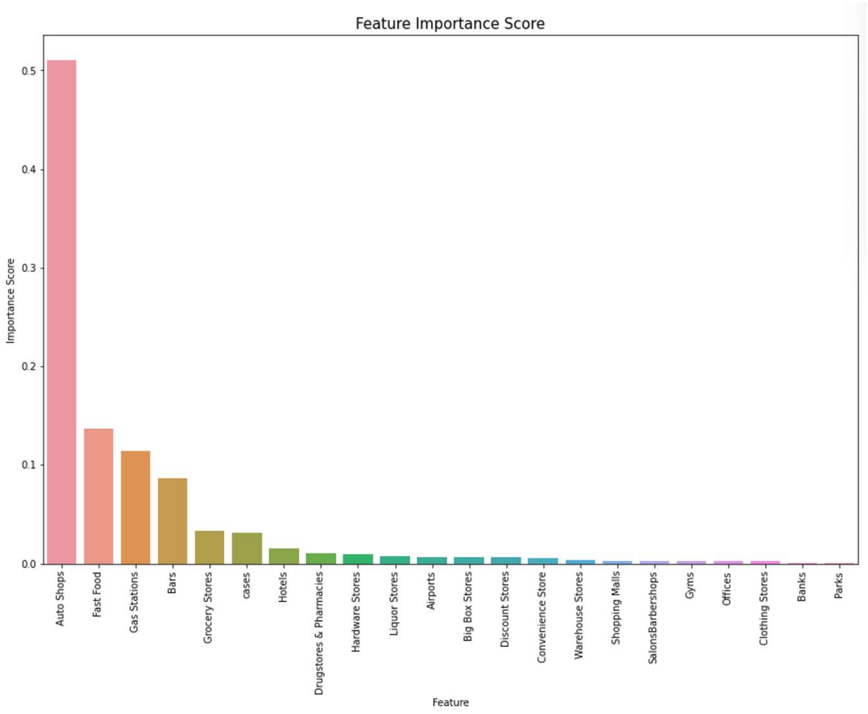


Figure 6.4 Feature importance

7 Impact and significance of the results

7.1 Significance of the results

The prediction model this project developed can predict the new future COVID-19 cases with an acceptable accuracy rate, especially when the target attributes are potentially of higher values.

The interactive tool, provided a user-friendly method for case prediction, as well as enabled the users to see the change of the new confirmed cases with varying foot traffic data instantly.

The features which are with higher correlation to the confirmed cases, can provide decision maker an alternative when considering shutting down all non-essential places and can potentially reduce the impact of this all-shut-down strategy.

7.2 Project limitations and future opportunities

Although this project's deliverables may have a very wide range of applications and a number of benefits, it is to keep in mind that it is a prototype and only based on data of one state in one country. To apply the applications to the wider world, more specific research and analysis must be carried out.

In addition, the datasets used in this project were lack of certain facets, especially the data collection details of the venues, despite the group's effort of obtaining such information from the data provider. As the transmission mechanism of this infectious disease is very complex, it is possible that many other factors would impact the transmission rate and the model can be impacted by uncertainties. Also, there is possibility that people get contracted do not get tested at the time that the incubation period ends. This can also impact the model performance in the reality.

In this model, the features can have positive as well as negative impact on the disease transmission. More research is required to analyse the reasons behind this and identify ways to differentiate these two types of features (e.g., using combined factors instead of individual features).

8 Project management

This project was delivered following an agile principle. Without an actual client, the agile methodology was adjusted. Instead of having regular client feedback, the team used the WIL project requirement, feedback from the tutor and the agreed questions to solve as the client instructions. As shown in Figure 8.1, during every one-week sprint development cycle, the project team assigned tasks, worked on tasks, made work ready for team review, and had discussions.

Figure 8.1 Project delivery methodology

The project started with an inception/ kick-off meeting to introduce team members and discuss the project objectives. During the project delivery period, a weekly meeting was held to get project progress update, review tasks completed, discuss issues identified and assign new tasks for the coming weeks. The regular and frequent catch-ups were shown to be helpful to bring every team member on the same page, to identify and mitigate risks in time, and keep the project on track.

8.1 Team organisational chart

The project team has a very diversified background and each team member has their own strength. After discussion, it was agreed that the team was to follow the below organisational chart as shown in Figure 8.2.

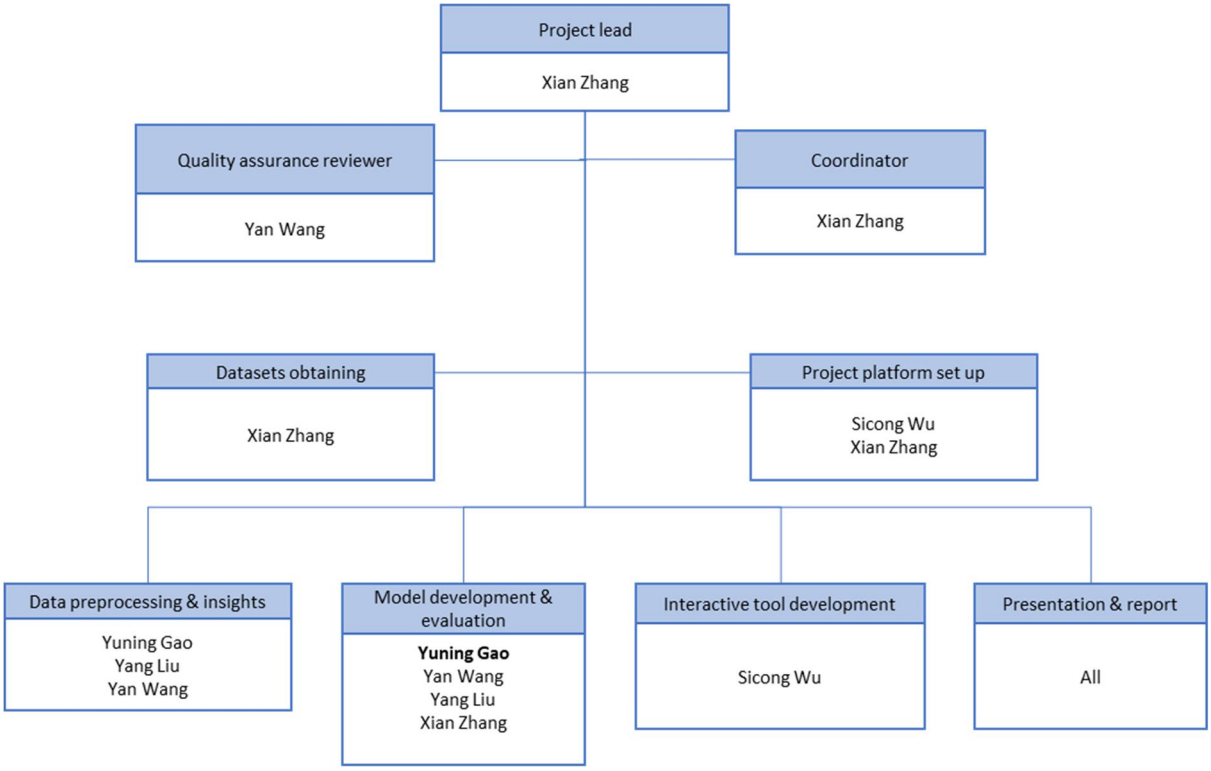


Figure 8.2 Team organisational chart

To ensure the work had good quality, a quality assurance scheme was applied, with each task completed to have at least one person assigned for review.

8.2 Management platform

Microsoft Teams was used as the primary platform for team communication and work exchange. A repository inside GitHub was created for code sharing.

8.3 Project program

The project was delivered following the below program as shown in Figure 8.3. A detailed program was also developed using Microsoft project and included in Appendix C.

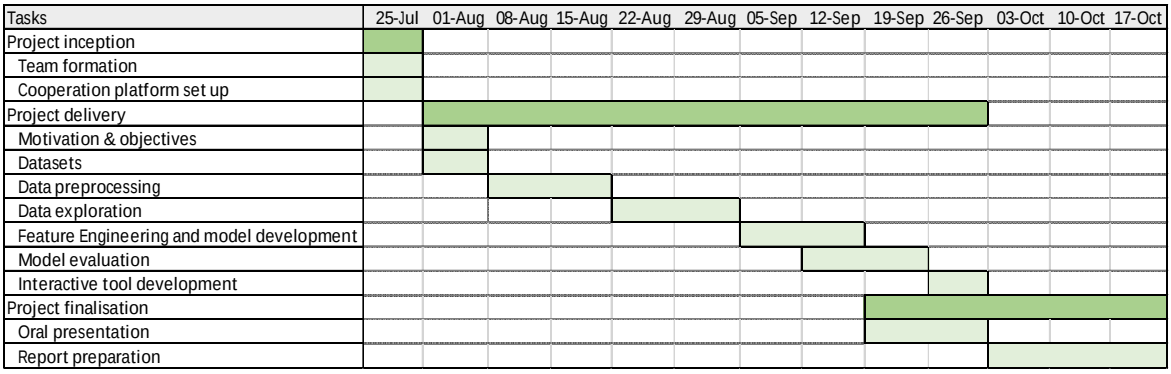


Figure 8.3 Project program (high-level)

8.4 Team contribution table

Each team member's contribution is shown in Table 1

Table 1 Contribution

Member	ID	Contribution	Main tasks
Yuning Gao	S3835212	20%	Model development
Yang Liu	S3289475	20%	Data insights & visualisation
Yang Wang	S3646791	20%	Data insights & project review
Sicong Wu	S3735283	20%	Interactive tool development
Xian Zhang	S3749428	20%	Project lead & modelling support

9 Reference

ABC news, 2020. *Stage 4 coronavirus restrictions in Melbourne could cost the economy $9 billion, Scott Morrison says.* [Online]
Available at: https://www.abc.net.au/news/2020-08-06/victoria-coronavirus-crisis-blows-out-gdp-estimates-by-billions/12530130
[Accessed 21 Aug 2020].
Australian Government Department of Health , 2020. *What you need to know about coronavirus (Covid-19).* [Online]
Available at: https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/what-you-need-to-know-about-coronavirus-covid-19
[Accessed 21 Aug 2020].
BBC news, 2020. *Coronavirus: A visual guide to the economic impact.* [Online]
Available at: https://www.bbc.com/news/business-51706225
[Accessed 21 Aug 2020].
Congressional Budget Office, 2020. *Interim Economic Projections for 2020 and 2021,* Washington : Congressional Budget Office.
Dickens, B. L. et al., 2020. *Modelling lockdown and exit strategies for Covid-19 in Singapore,* Singapore : The Lancet Regional Health .
Foursquare, 2020. *Covid-19 Foot Traffic Data.* [Online]
Available at: https://aws.amazon.com/marketplace/pp/prodview-cjhkgxpn6vcce?qid=1595561869554&sr=0-3&ref_=srh_res_product_title#overview
[Accessed 10 Aug 2020].

Health and Human Services Victoria, 2020. *Coronavirus Covid-19 in Victoria,* Melbourne : Health and Human Services Victoria.
International Monetary Fund, 2020. *Wolrd economic outlook update June 2020,* Washington: International Monetary Fund.
Kelly, P., 2020. *Victoria not alone in latest covid-19 response* [Interview] (4 July 2020).
Lauer, S. A., Grantz, K. H., Bi, Q. & Jones, F. K., 2020. *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application,* Philadelphia: Annals of Internal Medicine .
Medpage today, 2020. *As COVID Data Collection Moves From CDC to HHS, Questions Mount.* [Online]
Available at: https://www.medpagetoday.com/infectiousdisease/covid19/87632
[Accessed 10 October 2020].
Rearc, 2020. *Coronavirus(COVID-19) Data in the United States The New York Times.* [Online]
Available at: https://aws.amazon.com/marketplace/pp/prodview-jmb464qw2yg74?qid=1597313655129&sr=0-1&ref_=srh_res_product_title#overview
[Accessed 1 Aug 2020].
The New York Times, 2020. *New York Covid Map and Case Count.* [Online]
Available at: https://www.nytimes.com/interactive/2020/us/new-york-coronavirus-cases.html
[Accessed 10 October 2020].
U.S. Centers for Disease Control and Prevention , 2020. *Social distancing.* [Online]
Available at: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html
[Accessed 21 Aug 2020].
U.S. Centers for Disease Control and Prevention, 2020. *How Covid-19 spreads.* [Online]
Available at: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html
[Accessed 21 Aug 2020].
U.S. Centers for Disease Control and Prevention, 2020. *Symptoms of coronavirus.* [Online]
Available at: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
[Accessed 21 Aug 2020].
World Health Organization , 2020. *Q&A on coronaviruses.* [Online]
Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses
[Accessed 21 Aug 2020].
World Health Organization , 2020. *WHO Coronavirus Disease (Covid-19) Dashboard.* [Online]
Available at: https://covid19.who.int/
[Accessed 21 Aug 2020].
World Health Organization, 2020. *WHO Coronavirus Disease (COVID-19) Dashboard.* [Online]
Available at: https://covid19.who.int
[Accessed 18 10 2020].

Appendix A - data visualisation sample sheet

````
---
title: "R Notebook"
output:
  html_document:
    df_print: paged
runtime: shiny
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(shiny)
library(dplyr)
library(RColorBrewer)
library(plotly)
library(readr)
```

```{r}
DF <- read_csv("datadraw.csv",col_types = cols(place = col_factor(levels =
c('Airports',
 'Auto Shops',
 'Banks',
 'Bars',
 'Big Box Stores',
 'Clothing Stores',
 'Drugstores & Pharmacies',
 'Discount Stores',
 'Convenience Store',
 'Fast Food',
 'Gas Stations',
 'Grocery Stores',
 'Gyms',
 'Hardware Stores',
 'Hotels',
 'Liquor Stores',
 'Offices',
 'Parks',
 'Shopping Malls',
 'Warehouse Stores',
 'Est_New_Avg'
 ))))
DF1 <- DF %>% filter(place != "Est_New_Avg")
DF2 <- DF %>% filter(place == "Est_New_Avg")


accumulate_by <- function(dat, var) {
  var <- lazyeval::f_eval(var, dat)
  lvls <- plotly:::getLevels(var)
  dats <- lapply(seq_along(lvls), function(x) {
    cbind(dat[var %in% lvls[seq(1, x)], ], frame = lvls[[x]])
  })
  dplyr::bind_rows(dats)
}

DF2 <- DF2 %>% accumulate_by(~date)
```
````

Appendix B - model development coding sample sheet

Data Preprocessing

In [1]:

```
# Import basic packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
# Import USA Covid-19 daily cases
cases = pd.read_csv("covid-19-nyt-data-in-usa_dataset_us-counties.csv")
```

In [3]:

```
# Dataframe manipulation
ny_case = cases[cases['county'] == 'New York City']
ny_case = ny_case[["date", "cases", "deaths"]]
ny_case["new_cases"] = ny_case["cases"].diff()
ny_case["new_deaths"] = ny_case["deaths"].diff()
ny_case["date"] = pd.to_datetime(ny_case["date"])
ny_case = ny_case.set_index("date")
```

In [4]:

```
# Add new column for estimated new cases based on assumption
est_avg = []
for i in range(0, len(ny_case)):
    avg = int(ny_case["new_cases"][i+4: i+7].sum() / 3)
    est_avg.append(avg)
ny_case["Est_New_Avg"] = est_avg
```

In [5]:

```
# Check the dataframe index for concatenation
ny_case.index
```

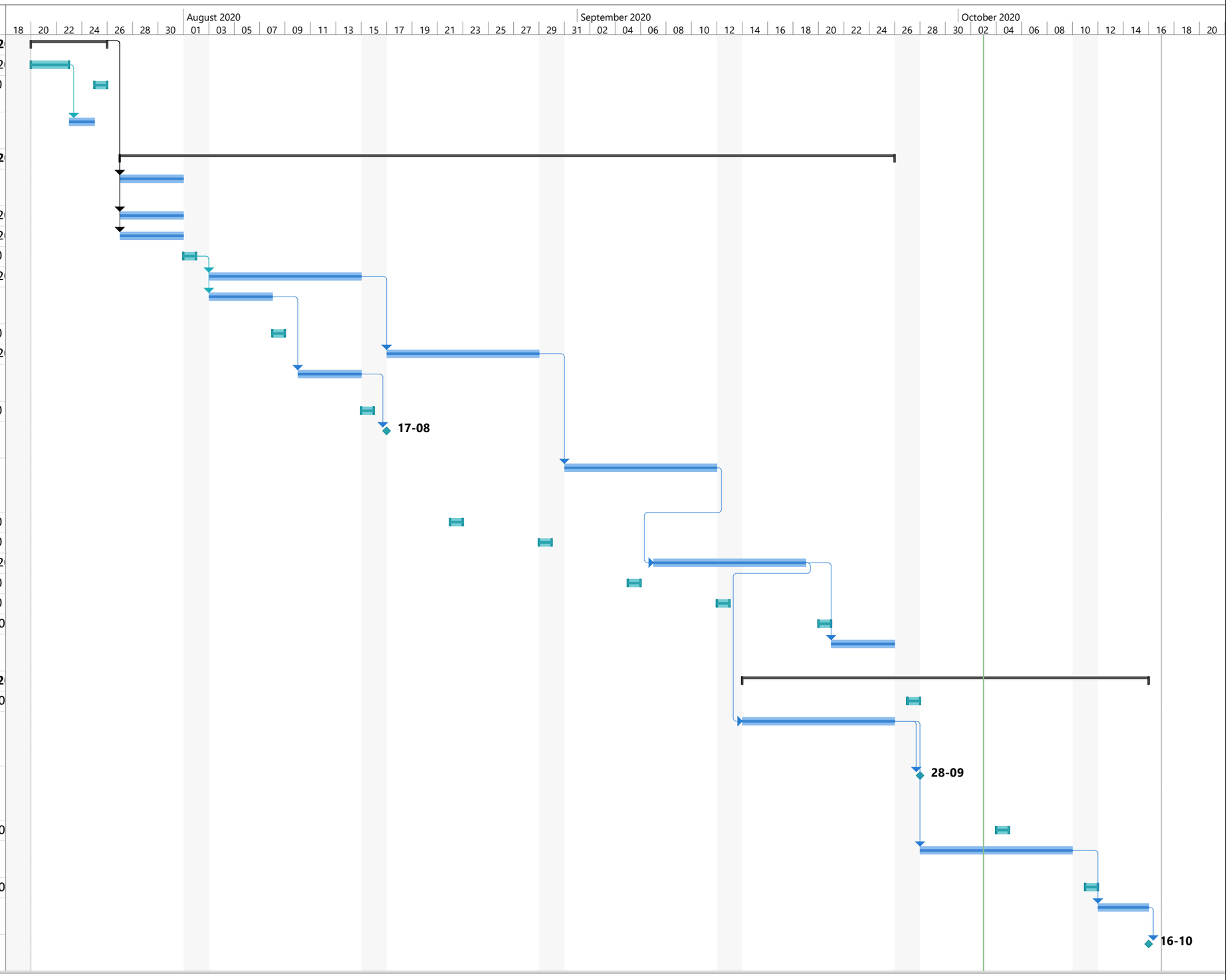
Out[5]:

```
DatetimeIndex(['2020-03-01', '2020-03-02', '2020-03-03', '2020-03-04',
               '2020-03-05', '2020-03-06', '2020-03-07', '2020-03-08',
               '2020-03-09', '2020-03-10',
               ...
               '2020-08-25', '2020-08-26', '2020-08-27', '2020-08-28',
               '2020-08-29', '2020-08-30', '2020-08-31', '2020-09-01',
               '2020-09-02', '2020-09-03'],
              dtype='datetime64[ns]', name='date', length=187, freq=No
ne)
```

Appendix C - detailed program

Project program

ID	Task Mode	Task Name	Duration	Start
1		**Project Inception Sta**	**5 days**	**Mon 20-07-2**
2		Team forming	3 days	Mon 20-07-2
3		Project inception meeting	1 day	Sat 25-07-20
4		Cooperation platform setup	2 days	Thu 23-07-20
5		**Project Delivery**	**45 days**	**Mon 27-07-2**
6		Motivation and question	5 days	Mon 27-07-20
7		Datasets selection	5 days	Mon 27-07-2
8		Team role assignm	5 days	Mon 27-07-2
9		Team meeting 01	1 day	Sat 01-08-20
10		**Data preprocessin**	10 days	Mon 03-08-2
11		WIL-milestone 1 preparation	5 days	Mon 03-08-20
12		Team meeting 02	1 day	Sat 08-08-20
13		**Data exploration**	10 days	Mon 17-08-2
14		Wil-milestone 1 review	5 days	Mon 10-08-20
15		Team meeting 03	1 day	Sat 15-08-20
16		Wil-milestone 1 submission	0 days	Mon 17-08-20
17		**Feature Engineering and model**	10 days	Mon 31-08-20
18		Team meeting 04	1 day	Sat 22-08-20
19		Team meeting 05	1 day	Sat 29-08-20
20		**Model evaluation**	10 days	Mon 07-09-2
21		Team meeting 06	1 day	Sat 05-09-20
22		Team meeting 07	1 day	Sat 12-09-20
23		Team meeting 08	1 day	Sun 20-09-20
24		**Interactive tool developent**	5 days	Mon 21-09-20
25		**Project finalisation**	**24 days**	**Mon 14-09-2**
26		Team meeting 09	1 day	Sun 27-09-20
27		WIL - Oral presentation preparation	10 days	Mon 14-09-20
28		WIL-Oral presentation submission	0 days	Mon 28-09-20
29		Team meeting 10	1 day	Sun 04-10-20
30		WIL Project - final report preparation	10 days	Mon 28-09-20
31		Team meeting 11	1 day	Sun 11-10-20
32		WIL Project -final report review	4 days	Mon 12-10-20
33		WIL Project - submission	0 days	Fri 16-10-20



Project: Project Program
Date: Sat 03-10-20

Task	Project Summary	Manual Task	Start-only	Deadline
Split	Inactive Task	Duration-only	Finish-only	Progress
Milestone	Inactive Milestone	Manual Summary Rollup	External Tasks	Manual Progress
Summary	Inactive Summary	Manual Summary	External Milestone	