Data analysis of COVID-19 based on different factors in USA

Xiaohan Ding[1]
Department of Computer Science
George Mason University
Fairfax, VA, US
xding2@gmu.edu.com

Cheng Zhong[1]
Department of Computer Science
George Mason University
Fairfax, VA, US
czhong2@gmu.edu.com

ABSTRACT

In 2020, as the impact of the epidemic continues to intensify, more social-level contradictions are intensified. Whether it is reinstatement or staying at home, there is an urgent need for a reasonable explanation and scientific supporting evidence. This project is focusing on the comprehensive exploration of COVID-19 based on various factors at this time and implement reliable analysis result and prediction model.

In this project, it is looking for the factors that could be important during the spread of COVID-19 and find out more information between those factors. The major research direction is based on USA spatiotemporal state-level data so we can minimize the policy difference from different states. Intuitively, the relatively general factors like population and medicine level have major influence, so this model is added more derived features. In this way, this research can build a model to predict future R0 and tendency based on these features.

The completion of this project has passed a total of 5 stages. In the first stage, the project team conducted a comprehensive understanding of the data, including: a description of the original data and EDA of the original data.

In the second stage, data pre-processing is started, which mainly includes: selecting appropriate data, data cleaning, filling in missing data, creating training and test data and generating the final data set csv format file.

In the third stage, the initial creation of the model includes: establishment of regression model, establishment of SIR model, integration of R0 with regression model and integration growth rate with regression model.

In the fourth stage is to carry out model optimization and prediction, which mainly includes: prediction of the test set, prediction for different target-labels, fitting curves and generating prediction values.

In the fifth stage, the model evaluation is completed, which mainly includes: calculating the loss function, calculating the valid-rmse, evaluating the accuracy of the prediction, and comparing and analyzing the weight and significance of various factors.

Our purpose is: We have to analyze how various policies in the United States affect the spread of COVID-19. Which kind of policy is the most powerful, and which kind of policy is the most urgently needed now. We will also complete the prediction of the epidemic, but this is just to support our policy analysis is correct.

KEYWORDS

Data mining, regression model, SIR model, COVID-19

1 Introduction

As COVID-19 spreads across the global, more and more attention from various research fields concentrates on related topics. From data analysis perspective, there are many available datasets collected by people from all walks of life, including basic epidemic statistic data (eg. confirmed numbers) and derived data which could be useful for related research. The transmission of the virus accelerated rapidly, which impacted billions of people in the world, also triggering the lockdowns in many countries. Under this situation, we are concerned about when the trend of pandemic will change, and what factors are playing an important role in turning around the situation. With these questions, we aim at the policy measures taken by the government that have been proved to be crucial in many other scenarios. For instance, school and business closure and public events cancellation helped China control the situation successfully, and many other policy measures are taking important parts [1]. First, we will process the existing data sets and merge them. Each time we analyze the data set, we will make a weight assumption, which may have different effects on COVID-19 propagation. The simple regression function we used at the beginning, that is, the characteristics of the weighted sum. When we try to analyze the accuracy of the basic model, the result of our test set is to reflect the change of R0. If R0 decreases, it means that our initial assumptions are correct. Then we can start the real optimization model.

We will input the test set into the model. This involves two test sets. The first is the inspection standard in New York last month where the government policy is not strict, and the second is the future case where the government policy is very strict. Through the test data, we can find that when the policy is not strict, the R0 of each state is very high, which means that more people will be infected. After strictly implementing this strategy, as the feature weight increases, the value of R0 decreases from 2 to 1.5. Therefore, a good policy strategy can control COVID.

With this novel idea, we implement the data mining model to explore the information behind policy datasets.

2 Related Work and Literature review

The spread of COVID-19 is influenced and even determined by many related factors such as policies [2], environment [2] and activities of people [2]. Among these factors, some are highly associated with others and some are relatively more independent. For instance, policies are affected by regional factors like people's common habits, working hours and so forth. To minimize the influence from unused associated factors, we define our data area within the US, which has considerable number of states to do experiments. Given the epidemic datasets, the SIR epidemic model [3] is suitable for COVID-19 scenario that explains the abstract process of virus spreading. In other cases, the SIR model can also be used to determine the late stage of an epidemic, when active cases number converge to zero [7]. In time-delayed datasets, it has also been proved reliable [4].
In recently research, we summarize some articles.
Firstly, regarding the mathematical study of infected individuals, its definition is that in the susceptible population, the expected number of secondary cases produced during the infection period is the main eigenvalue of the positive linear operator. The results of the research indicate that this feature value can usually be calculated or estimated easily. The article presents some examples involving various structural variables, such as people movement and activity tendencies[5].

The second case study focuses on re-infected infectious diseases, such as whether people who have recovered have the possibility of re-infection. Their experiment examined the relationship between the prevalence of repeated infections and the basic reproductive number (R0)[13]. This is also a good experiment to find the cause, which can bring a lot of inspiration to our research. First, the study resolved the general deterministic compartment model of reinfection to derive an analytical solution to this relationship. Then, we numerically solved a disease-specific model that explicitly tracks the spread of reinfected syphilis[13].

The last study. The article preached that they used the model to create the EVIDENCEMINER system[14], which is an automatic text mining system. We are curious about how this system works and how to use it to serve the epidemic. EVIDENCEMINER is a web-based system that enables users to query natural language statements and automatically retrieve textual evidence of life sciences from a background corpus. It is built in a completely automated way, without any manual training data annotation[14].

3 EDA and Data preprocessing

Our dataset mainly comes from four aspects:
The first is real-time updated time series data, and the core is the number of patients: https://github.com/stccenter/COVID-19-Data
The second is the policy and management data of various states in the United States: https://github.com/CSSEGISandData/COVID-19.git
The third is visual data updated in real time: https://npgeo-corona-npgeo-de.hub.arcgis.com/
The fourth is the weather data, which is only used to assist our project, not the core of our research: https://darksky.net/

3.1 Data description and size of data

First of all, our first data set was screened for data: In the four core data sets of *time_series_covid19_confirmed_global.csv*, *time_series_covid19_confirmed_US.csv*,*time_series_covid19_deaths_global.csv* , and *time_series_covid19_deaths_US.csv*, we first explore the meaning behind the data.

	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	...	5/5/20	5/6/20	5/7/20	5/8/20	5/9/20	5/10/20	5/11/20
0	16	AS	ASM	16	60.0	NaN	American Samoa	US	-14.2710	-170.1320	...	0	0	0	0	0	0	0
1	316	GU	GUM	316	66.0	NaN	Guam	US	13.4443	144.7937	...	145	149	149	151	151	151	151
2	580	MP	MNP	580	69.0	NaN	Northern Mariana Islands	US	15.0979	145.6739	...	14	15	15	15	16	16	19
3	630	PR	PRI	630	72.0	NaN	Puerto Rico	US	18.2208	-66.5901	...	1924	1968	2031	2156	2173	2198	2256
4	850	VI	VIR	850	78.0	NaN	Virgin Islands	US	18.3358	-64.8963	...	66	66	66	68	68	69	69

Figure.1 Sample data

And, we conduct a preliminary analysis of the data format.
In the **time_series_covid19_confirmed_US**, it contains:
RangeIndex: 3261 entries,
0 to 3260, Columns: 125 entries, UID to 5/17/20,
dtypes: float64(3), int64(116), object(6),
memory usage: 3.1+ MB.

In the **time_series_covid19_confirmed_global**, it contains:
RangeIndex: 266 entries,
0 to 265,Columns: 118 entries,
Province/State to 5/14/20,
dtypes: float64(2), int64(114),
object(2),memory usage: 245.3+ KB

In the **time_series_covid19_deaths_US**, it contains:
RangeIndex: 3261 entries, 0 to 3260
Columns: 126 entries, UID to 5/14/20
dtypes: float64(3), int64(117), object(6)
memory usage: 3.1+ MB

In the **time_series_covid19_deaths_global,** it contains:
RangeIndex: 266 entries, 0 to 265
Columns: 118 entries, Province/State to 5/14/20
dtypes: float64(2), int64(114), object(2)
memory usage: 245.3+ KB

In the US_State_Policy_2020-05-14 dataset, it contains the main features as follows,
['StateName', 'StateCode', 'S1_School Closure', 'S1_IsGeneral', 'S1_Notes', 'Index_School_Closure', 'S2_Workplace closing', 'S2_IsGeneral', 'S2_Notes', 'Index_Workpalce_Closure', 'S3_Cancel public events', 'S3_IsGeneral', 'S3_Notes',

'Index_Cancel_Public_Events', 'S4_Close public transport', 'S4_IsGeneral', 'S4_Notes', 'Index_Close_Public_Transport', 'S5_Public information campaigns', 'S5_IsGeneral', 'S5_Notes', 'Index_Public_Information_Campaigns', 'S6_Restrictions on internal movement', 'S6_IsGeneral', 'S6_Notes', 'Index_Restrictions_On_Internal_Movement', 'S7_InternationalNational travel controls', 'S7_Notes', 'Index_International_National_travel_controls', 'S8_Fiscal measures', 'S8_Notes', 'S9_Monetary measures', 'S9_Notes', 'S10_Emergency investment in health care', 'S10_Notes', 'S11_Investment in Vaccines', 'StringencyIndex'].

3.2 Exploratory analysis of raw data

Based on several datasets, we make a simple visualization of some important states which belongs to US, and find the tendency of growth rate.

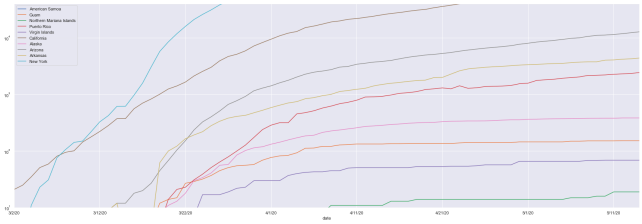

Figure.2 US Interstate data growth graph

We found that if we use the number of patients diagnosed as our indicator, it will not be a rational value, and the label value will change more and more, and we cannot get a stable prediction. Because our project aims at the analysis of different policies, reliable prediction results can well support our conclusions. So first of all, we conducted a log function transformation on the confirmation number to explore its rationality.

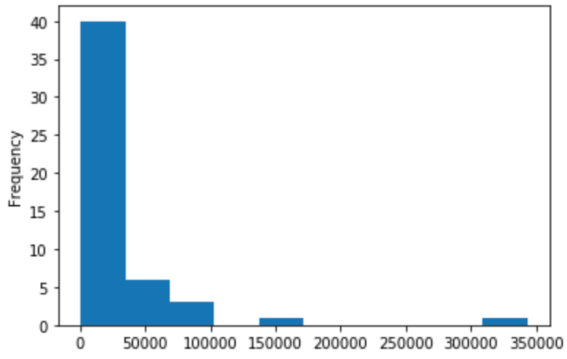

Figure.3 Distribution of raw data

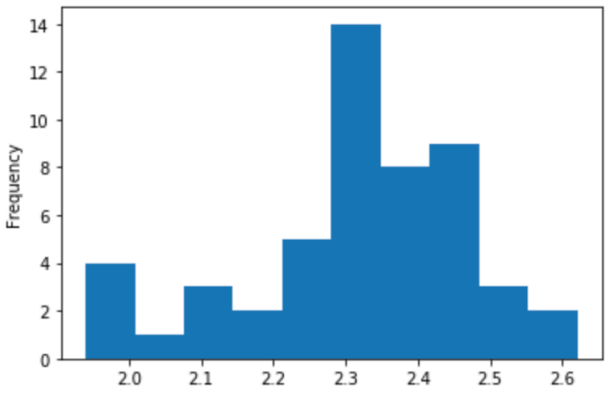

Figure.4 log (raw data) distribution

The data set label becomes our initial processing strategy after being transformed by the log function. It shows that this is not the best, but in the early stage of the project, we tested according to the value of this label. In the following text, I will explain in detail if we optimize the label step by step, and finally determine to use R0 as our Label value.

3.3 Data preprocessing for serval estimation

Now, we start our data preprocessing operation according to the most preliminary plan. The huge amount of data makes us not very good at analyzing the growth trend, so we calculated the slope of growth separately according to the following formula and added it to our data set.

$$x \to log(x) \tag{1}$$

And then, for the conversion data we have obtained, obtain the slope or the rate of data growth.

$$df(x) \to d(log_a x) = (1/xlna)dx \tag{2}$$

After completing all the operations on tags, we started to filter the features that are not needed for various data sets, which mainly not include the following features in the policy data set and we remain them to help us finish analysis:
['State',
'confirmed',
'Index_School_Closure',
'Index_Workpalce_Closure',
'Index_Cancel_Public_Events',
'Index_Close_Public_Transport',
'Index_Public_Information_Campaigns',
'Index_Restrictions_On_Internal_Movement',
'Index_International_National_travel_controls',
'StringencyIndex']

In order to better merge the features, we will weight the feature values and analyze the policy data[12] to be better than before.

Data analysis of COVID-19 based on different factors in USA

We have listed the meaning of each data. This is also a process of estimation.

Attribute Name	Description	Example
Data	Date	20200301
State Name	Name of the state	Alabama
StateCode	State name abbreviation	AL
S1_School Closure	Ordinary scale records closings of schools and universities	0- No measure 1- Recommend closing 2- Require closing
S1_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S1_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_School_Closure	Calculated weight for school closure	0, 33.33, 66.66, 100
S2_Workplace closing	Ordinary scale records closings of workplaces	0 – No measure 1- Recommend closing 2- Required closing
S2_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S2_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
S3_Cancel public events	Ordinary scale records canceling public event	0 – No measure 1- Recommend cancelling 2- Required cancelling
S3_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S3_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_Cancel_Public_Events	Calculated weight for canceling public events	0, 33.33, 66.66, 100
S4_Close public transport	Ordinary scale records closings of public transport	0- No measure 1- Recommend closing 2- Required closing
S4_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S4_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_Close_Public_Transport	Calculated weight for canceling public transport	0, 33.33, 66.66, 100
S5_Public information campaigns	Ordinary scale records presence of public info campaigns	0- No COVID-19 public information campaign 1- COVID-19 public information campaign
S5_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S5_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_Public_Information_Campaigns	Calculated weight for public information campaigns	0, 33.33, 66.66, 100
S6_Restrictions on internal movement	Ordinary scale records closings restrictions on internal movement	0- No measure 1- Recommend closing 2- Required closing
S6_IsGeneral	Binary scale for geographic scope	0 – Targeted 1 - General
S6_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_Restrictions_On_Internal_Movement	Calculated weight for restricting internal movement	0, 33.33, 66.66, 100
S7_InternationalNational travel controls	Ordinary scale records restrictions on international/national travel	0- No measure 1- Screening 2- Quarantine on high-risk regions 3- Ban on high-risk regions
S7_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
Index_International_National_Travel_controls	Calculated weight for canceling public transport	0, 33.33, 66.66, 100
S8_Fiscal measures	What economic stimulus policies are adopted?	Encouraged businesses to implement sick leave policy so sick employees, do not come to work and expose others because of financial concerns.
S8_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
S9_Monetary measures	What monetary policy interventions?	1: Three-Way Agreement with Legislature on Paid Sick Leave Bill to Provide Immediate Assistance for New Yorkers Impacted By COVID-19
S9_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
S10_Emergency investment in health care	Short-term spending on, e.g, hospitals, masks, etc	Governor Larry Hogan today issued a directive requiring all state health insurers to waive costs associated with testing for COVID-19. The directive, issued under the governor's authority during a state of emergency, waives any cost-sharing, including co-payments, coinsurance, and deductibles, in order to remove cost barriers to testing.
S10_Notes	Website link for the specific policy	https://www.mass.gov/doc/march-16-2020-k-12-school-closing-order
S11_Investment in Vaccines	Announced public spending on vaccine development	
StringencyIndex	The calculated stringency index according to the policy	14.28571429

Figure.5 Attribute Name with Description

Next, we remove some outliers and fill in missing values.

ALG_QUANTILE_CLIP(*group*)

If (group[group < group.quantile(.a)])
  group[group < group.quantile(.a)] ←group.quantile(.a)
ELSE
  group[group > group.quantile(1-a)] ← group.quantile(1-a)
  RETURN *group*

And then we finish the first step of preprocessed

3.4 Data preprocessing for filtering and merging

Before filtering and merging, the basic thing is to check whether there are missing values in the data set, and fill in the missing values. After this, we can filter and merge the data set.

ALG_DATE_TREND(group)

```
tmp = group.groupby('feature1').mean().reset_index()
def nan_helper(y):
    return np.isnan(y), lambda z: z.nonzero()[0]

y = tmp['feature2'].values
nans, x = nan_helper(y)
if group.link_ID.values[0] in RANGE
    tmp['date_trend'] = group['feature2'].median()
else:
    regr = linear_model.LINEARREGRESSION()
    regr.fit(x(~nans).reshape(-1, 1), y[~nans].reshape(-1, 1)
group = pd.merge(group, tmp[['feature1', 'feature2']],
on='date_hour', how='left')
return group
```

And using **LINEARREGRESSION** in **ALG_DATE_TREND**(), the data has been supplemented with missing values[12], the missing values will be filled according to the trend calculated above. In this way, the data will not be unstable because of the average. Using the trend of data, we can calculate the future

situation of the feature well. For example, New York 's government policy is a good example. At the beginning, New York did not pay attention to the policy of "company closure", which led to this characteristic data has been promoting a large increase in the number of confirmations. Then, when the weight of the policy of 'company closure' increased, it effectively suppressed the growth rate of the epidemic. This is a good reflection of the growth trend of policy, which can affect the growth trend of the number of patients with outbreaks.

Finally, we use the id number of each state to merge the data set, *pd.merge(dt1, dt2 _train, on=['State_id'], how='left')*.
And then we will get the final training set:

	State	confirmed	Index_School_Closure	Index_Workpalce_Closure	Index_Cancel_Public_Events	Index_Close_Public_Transport
0	New York	343051	100.0	100.000000	100.000000	66.666667
1	New Jersey	141902	100.0	66.666667	100.000000	66.666667
2	Illinois	87786	100.0	100.000000	100.000000	66.666667
3	Massachusetts	81882	100.0	100.000000	100.000000	0.000000
4	California	74861	100.0	66.666667	66.666667	66.666667

Figure.6 training set

4 Infectious Disease Model

Our infectious disease model is not the core of our project, but it is a good process to explore the changing trends of the number of patients in depth.
If we directly deploy directly according to the model, this is not a good strategy. Deploying the model also needs to be aware of what the model's assumptions are, how far it may be from reality, and which parameters are included. In general, it is the famous saying of George Box: Compared with the complexity of reality, all models Wrong, but some models are useful.[6]

4.1 SIR Model

The classic SIR model is a classic infectious disease model invented in the early last century. This model can roughly show the process from the onset to the end of an infectious disease[7].

S: is a susceptible group
I: is infected
R: is the recovery crowd
infection rate: β
recovery rate: γ

These three quantities are all functions that follow the change of time, which can be expressed as, where t is set to a unit time, we have the following formula[7]:

$$\frac{ds(t)}{dt} = -\beta s(t) I(t) \quad (3)$$

$$\frac{di(t)}{dt} = \beta S(t) I(t) - \gamma i(t) \quad (4)$$

$$\frac{dr(t)}{dt} = \gamma i(t) \quad (5)$$

And then running this algorithm.

SIR_MODEL(SIR, β, γ):
S, I, R←SIR
*S_ ←-β*S*I/N0
*I_ ←β*S*I/N0- γ *I
*R_← γ *I
return(*S_ , *I_ , *R_)

This model is a one-way model. The number of susceptible people is constantly input to the number of infected people, and at the same time the number of infected people is also input to the number of recovered people in one direction.

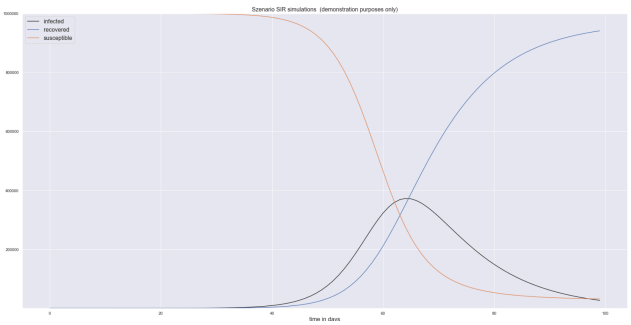

Figure.7 SIR

Everyone will be the number of restorers, which is the limitation of this model.

4.2 Fatality rate, R0

First look at the case fatality rate. We think the more appropriate term should be Case Fatality Ratio[11]. The biggest difference between Rate and Ratio is that Rate often introduces the time dimension in the denominator.
This measurement tool is very intuitive, that is, in a certain period of time, the proportion of patients who die of a certain disease (denominator), and therefore die (numerator). The problem is that this definition does not directly relate to how long a certain period of time is. The most ambiguous part of this definition is here[10]. Ideally, if the onset of a certain disease is cured or the time to death is fast, we can judge the recovery and death of all diagnosed patients in a short period of time, so as to obtain a more accurate estimate of the mortality rate. However, the occurrence of an epidemic is an uninterrupted process. There are new cases and patients with a long course of disease who have not recovered or died. How to estimate the mortality rate in this process is a problem. From the perspective of numerator and denominator.
The denominator is a patient who has been diagnosed. Under the condition of limited testing, this number must be less than the current number of infected people. Among them, the proportion of critically ill patients is higher (the false yin and false yang conditions for testing are not considered here).[10] Under such circumstances as the infection in nursing homes in Washington State, most of the denominators have basic diseases at an advanced

age. From these perspectives, the current estimated case fatality rate may overestimate the case fatality rate among a wider population.

R0 is an indicator of pathogenic contagiousness. It estimates an infectious infected person at the beginning of an outbreak of infectious disease (when no one is immune), how many people can be directly infected[5].
For example, as shown in the figure below, gray is the infected person, and white is the susceptible person who is not infected with the pathogen. From this estimate, R0 is probably about 2[6].

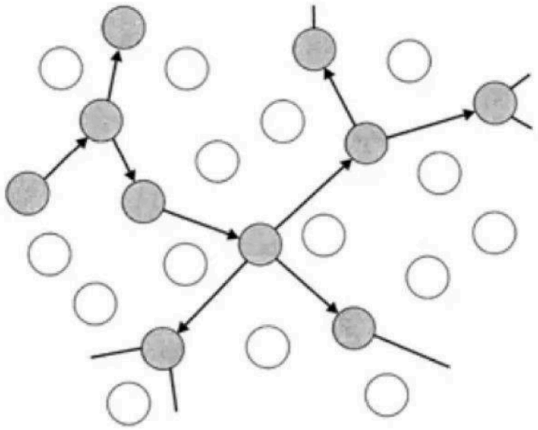

Figure.8 Sample of R0

Under this condition, we will take R0 be the new label value in our training set, and then we will implement this algorithm to finish calculation the R0. Definition of basic reproductive number (R0): At the beginning of the disease, when all people are susceptible, the number of people infected by a patient during their average illness period The R0 value formula calculated based on the SEIR model is as follows[8]:

$$R0 = 1 + \lambda T_g + \rho(1-\rho)\left(\lambda T_g\right)^2 \quad (6)$$

$$\lambda = lnY(t)/t \quad (7)$$

ALG_R0func(defi,susp,t)[8]

```
defi ← number of diagnoses;
susp ←the number of suspects;
t ← the time the disease has broken
p = 0.695
Tg_1 = 8.4
Tg_2 = 10.0
yt ← the actual estimated number of infected people
lamda = math.log(yt)/t

R0_1 = 1 + lamda * Tg_1 + p * (1 - p) * pow(lamda * Tg_1,2)
R0_2 = 1 + lamda * Tg_2 + p * (1 - p) * pow(lamda * Tg_2,2)

RETURN R0_1,R0_2
```

5 Regression model with Policy Analysis

The updated template, user manuals, samples, and required fonts, all are available at the URL https://www.acm.org/publications/proceedings-template. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template,

Firstly, we created a hypothesis. After filtering the features in the policy data set, we got a completed data set. We assume that this is the domain**.**

$$\boldsymbol{POLICYName_{SET}} = \{X_{name} | x_{name} \in X,\ each\ x_i\ is\ [\\ 'SchoolClosure', \\ 'WorkpalceClosure', \\ 'CancelPublicEvents', \\ 'ClosePublicTransport', \\ 'PublicInformationCampaigns', \ldots\ldots']\}$$

$$|\{\boldsymbol{POLICYName_{SET}}\}| = 11 \quad (8)$$

Then normalize the data in the data set, it is important to use the weight to represent the core strength of each policy

$$\boldsymbol{POLICY_{SET}} = x\{X_i | x_i \in X_i,\ \ x_i\ range\ is\ [0,33,66,100]\}$$

We assume a function and get a regression model through regression analysis. Let the FEATURE_SET becomes {X1, X2,…Xn}, From common example:

$$\textbf{FEATURESET} = \{\mathbf{x}|\boldsymbol{x1} = \text{'Index\_School\_Closure'}, \boldsymbol{x2} = \text{'Index\_Workpalce\_Closure'},\ldots\} \quad (9)$$

After the regression:

$$\boldsymbol{result} = \{\boldsymbol{f(Xi = i)}|\boldsymbol{f}(\Theta 1 * \text{w(Index\_School\_Closure')} \ \&\ \Theta 2 * \text{w('Index\_Workpalce\_Closure'))} = \text{label value(R0)}\} \quad (10)$$

So, in our project, we have several features in POLICY_SET and COVID-19_SET, find the optimal parameter to fit

$$\text{from } x \text{ to be Xi} \rightarrow f(\text{Xi}) \quad (11)$$

After combination or integration: (more than two features), and add the error rate we will implement with such ALG based on xgboost.

ALG(train, test, params, features, n)

```
CREATE DATAFRAME
    train←pd.DataFrame
    'id, ←zipcode
    'lable' ← 'R0' or 'increasing rate'
    'pred'←np.zeros
```

TEST SUBMISSION RESULTS
test_pred with id
CROSS-VALIDATION
kfolder ← KFold(n_splits=nfold)
CONSTRUCTION TEST DMATRIX
tst ← DMatrix(data=test[feature_names])
for each (fold_id)
TRAVERSE EACH FOLD OF DATA IN CV,
CONSTRUCT THE CURRENTLY VALIDATED DMATRIX
xgb_val = xgb.DMatrix(
train.iloc[val_idx][feature_names],
train.iloc[val_idx]['confirmed'])
TRAINING A REGRESSION MODEL
reg ← train(params=params, dtrain=xgb_trn, **fit_params)
print(nCV LOSS, RMSE)
return test_pred
RETURN verification results

5.1 Preliminary analysis based on train/test set

First of all, through the regression model, the deployment of the model is completed normally, as long as the feedback data includes:
Will train until valid-rmse hasn't improved in 360 rounds.
[300] train-rmse:0.02389 valid-rmse:0.024229
[600] train-rmse:0.006856 valid-rmse:0.010381
[900] train-rmse:0.00496 valid-rmse:0.010253
[1200] train-rmse:0.003777 valid-rmse:0.010266
Stopping. Best iteration:
[1022] train-rmse:0.004413 valid-rmse:0.010229
CV LOSS: 0.000129050223541496

```
[0]     train-rmse:0.420253     valid-rmse:0.419024
Multiple eval metrics have been passed: 'valid-rmse' will be used for early stopping.

Will train until valid-rmse hasn't improved in 360 rounds.
[300]   train-rmse:0.023808     valid-rmse:0.024696
[600]   train-rmse:0.006483     valid-rmse:0.010892
[900]   train-rmse:0.004668     valid-rmse:0.010741
[1200]  train-rmse:0.003535     valid-rmse:0.010781
Stopping. Best iteration:
[913]   train-rmse:0.004603     valid-rmse:0.010731

[0]     train-rmse:0.419486     valid-rmse:0.42207
Multiple eval metrics have been passed: 'valid-rmse' will be used for early stopping.

Will train until valid-rmse hasn't improved in 360 rounds.
[300]   train-rmse:0.02389      valid-rmse:0.024229
[600]   train-rmse:0.006856     valid-rmse:0.010381
[900]   train-rmse:0.00496      valid-rmse:0.010253
[1200]  train-rmse:0.003777     valid-rmse:0.010266
Stopping. Best iteration:
[1022]  train-rmse:0.004413     valid-rmse:0.010229


CV LOSS: 0.000129050223541496
```

Figure.9 training output

Next, we put the test set into the model for deployment. This involves two test sets. The first is the test set in New York in January and February when the government policy is not strict, and the other is the recent case where the government policy is very strict (that is, the weight of the feature data is high).

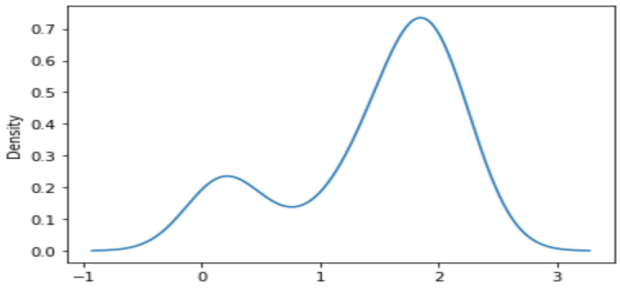

Figure.10 NY January and February test set

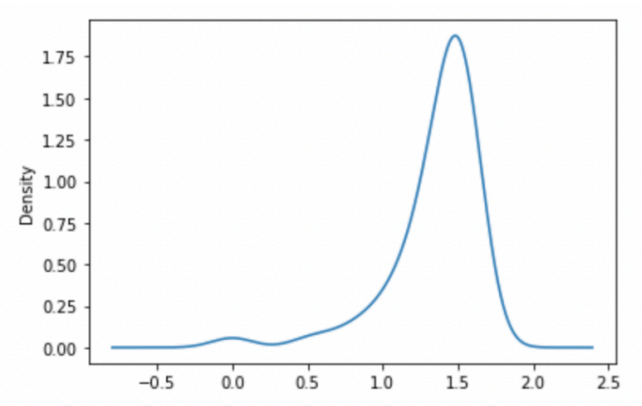

Figure.10 recent NY test set

Through the test data trend in the figure, we can well find that when the policy is not strict, the R0 in New York is very high, which means that more people will be infected. And after the policy is strict, as the weight of feature data continues to increase, the concentration point of R0 has been reduced from 2 to around 1.

Unfortunately, even though R0 is declining drastically, the overall density is on the rise. The reason for the increase in density is because the total number of patients is still in a large growth situation.

$$|Set1(sum\ of\ patient\ in\ MAY)| \gg |Set2(sum\ of\ patient\ in\ APRIL)|$$

5.2 In-depth analysis of the impact factor intensity based on PDP

PDPbox should be based on the stable model. Then, it contains serveral functions[9]:

```
from matplotlib import pyplot as plt
from pdpbox import pdp, get_dataset, info_plots
```

1.Auxiliary functions are used to visualize target distribution and forecast distribution.
2.The correct way to handle the one-key encoding function.
3.Solutions that solve complex interdependencies between functions.

4.Support multiple classifiers.
5.Local dependency graph supporting the interaction of two variables.

Next, we calculate how high each factor affects R0 according to the PDP graph. So we separate each feature.

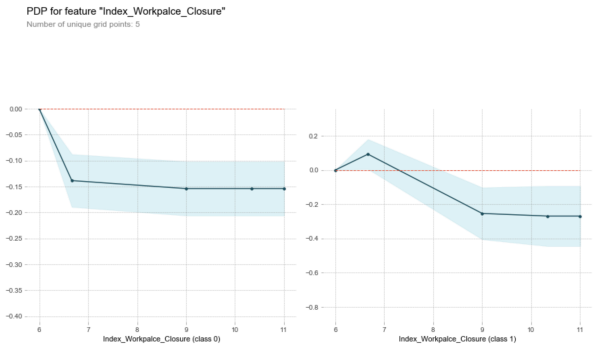

Figure.11 pdp analysis with R0

By calculating the distance between the blue curve and the red curve of R0, we can get the maximum impact strength. Here we can use Manhattan distance or Euclidean distance for calculation. If the blue curve appears below R0, it means that this will greatly reduce the spread of the epidemic. If the blue curve is above R0, these positive numbers mean that the current weight of this feature will greatly facilitate the spread of the epidemic. After the calculation is complete, reorder each feature. So we can get a sorted array of impact factors.

Weight	Feature
0.3385 ± 0.1231	confirmed
0.0462 ± 0.1231	Index_Workpalce_Closure
0.0308 ± 0.0754	StringencyIndex
0 ± 0.0000	Index_Restrictions_On_Internal_Movement
0 ± 0.0000	Index_Public_Information_Campaigns
0 ± 0.0000	Index_School_Closure
-0.0154 ± 0.0615	Index_Close_Public_Transport
-0.0154 ± 0.0615	Index_Cancel_Public_Events
-0.0308 ± 0.1231	Index_International_National_travel_controls

Figure.11 sorted array of impact factors.

Here is different states analysis with different class:

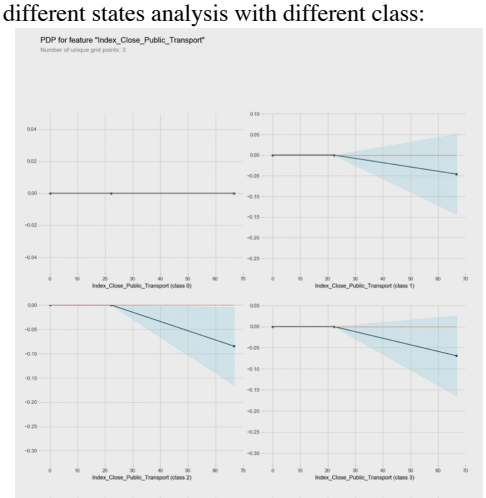

Figure.12 analysis for different class

6 Result and Contribution

In the visualization of pdp, if the blue curve appears below R0, this means that this will greatly reduce the spread of the epidemic. If the blue curve is above R0, these positive numbers mean that the current weight of this feature will greatly facilitate the spread of the epidemic. What we want is to achieve the best control of the most epidemic situation through the distribution of a weight. So in the end, Investigation of the learned model with the important features are 'workplace_close' and 'StringencyIndex'. After that, we can finally do the final parameter optimization and assign different weights to different features in an orderly manner. Then make the last forecast. Finally, we use the evaluation of the model to test the accuracy of the prediction:
[0.70790123, 0.69550265, 0.65987934, 0.73361017, 0.81982478]
[0.723343634]

ALG_ EVALUATION()

```
ad = pd.read_csv()
ad['actual_label'] = ad ['admit']
ad = ad.drop()

kf = KFold(len(ad), 5, shuffle, random_state)
ir = logisticregression()

accuracy = cross_val_score(ir, ad['label name'], cv=kf)
average_ accuracy = sum(ad)/len(ad)
RETURN average_ accuracy
```

This prediction is aimed at what kind of impact it will have if the United States chooses to resume work.
If the United States chooses to resume work, we need to subtract an average value for each feature value in test set.

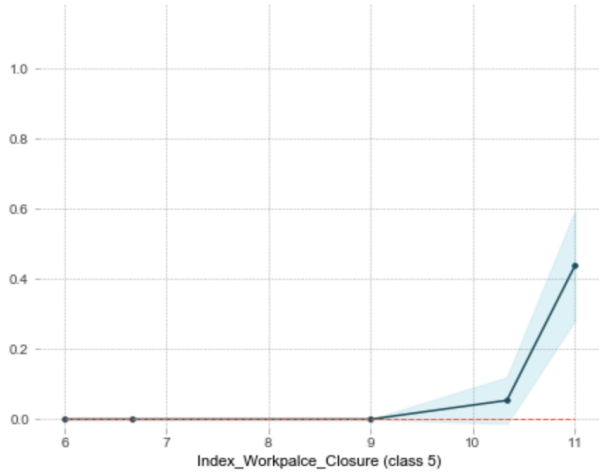

Figure.13 test set for work is resumed

After the reworked test set was deployed in the model, we found that this greatly increased the propagation rate, and R0 was also positively affected and began to rise rapidly. This is an undesirable choice. The epidemic will continue to grow, and there is no possibility of a decline in June.

Data analysis of COVID-19 based on different factors in USA

CONTRIBUTION

Xiaohan Ding:
Patient data collection and preprocessing, regression model and policy analysis, model optimization

Cheng zhong:
Government data collection and preprocessing, R0 and infectious disease models, model evaluation

REFERENCES

[1] Wang, J., Tang, K., Feng, K. and Lv, W., 2020. High temperature and high humidity reduce the transmission of COVID-19. Available at SSRN 3551767.

[2] Shereen, M.A., Khan, S., Kazmi, A., Bashir, N. and Siddique, R., 2020. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research.

[3] McCluskey, C.C., 2010. Complete global stability for an SIR epidemic model with delay—distributed or discrete. Nonlinear Analysis: Real World Applications, 11(1), pp.55-59.

[4] Beretta, E. and Takeuchi, Y., 1995. Global stability of an SIR epidemic model with time delays. Journal of mathematical biology, 33(3), pp.250-260.

[5] Diekmann, O. , Heesterbeek, J. A. P. , & Metz, J. A. J. . (1990). On the definition and the computation of the basic reproduction ratio r0 in models for infectious-diseases in heterogeneous populations. Journal of Mathematical Biology, 28(4), 365-382.

[6] Chuan-Qing, X. U. , Xiao-Xiao, W. , Jing-An, C. , Xiao-Jing, W. , & Science, S. O. . (2019). Infectious disease model and optimal immune research about the influence of external population on guangdong province tuberculosis. mathematics in practice and theory.

[7] Shulgin, B., Stone, L. and Agur, Z., 1998. Pulse vaccination strategy in the SIR epidemic model. Bulletin of mathematical biology, 60(6), pp.1123-1148.

[8] Peeters, K. C. M. J. , Kattan, M. W. , Hartgrink, H. H. , Kranenbarg, E. K. , Karpeh, M. S. , & Brennan, M. F. , et al. (2005). Validation of a nomogram for predicting disease-specific survival after an r0 resection for gastric carcinoma. Cancer, 103(4), 702-707.

[9] Gao, W. , Sanna, M. , & Wen, C. P. . (0). Geo-temporal distribution of 1,688 chinese healthcare workers infected with covid-19 in severe conditions – a secondary data analysis. Social ence Electronic Publishing.

[10] A, Y. N. M. , A, T. T. H. , A, J. X. Z. , A, Q. Q. , A, Y. X. G. , & A, S. Y. L. , et al. (2020). Estimating instant case fatality rate of covid-19 in china. International Journal of Infectious Diseases.

[11] Angelopoulos, A. N. , Pathak, R. , Varma, R. , & Jordan, M. I. . (2020). On the bias arising from relative time lag in covid-19 case fatality rate estimation.

[12] Wang, H. , Zhu, R. , & Ma, P. . (2019). Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 114(525), 486-486.

[13] Feldman, J. , & Mishra, S. . (2019). What could re-infection tell us about r0? a modeling case-study of syphilis transmission. Infectious Disease Modelling, 4, 257-264.

[14] Wang, X. , Liu, W. , Chauhan, A. , Guan, Y. , & Han, J. . (2020). Automatic textual evidence mining in covid-19 literature.