COVID-19 Forecasting And Analyzation

Weibang He, Yilin Zhou, Shuqing Zhang, Yuan Wan
Computer Engineering Department
San Jose State University, San Jose, California, USA

***Abstract*— In response to the COVID-19 pandemic, people from all kinds of science domain and industry domain work together and try their best to figure out the problem faced by all over the world. Global research has been taken to apply recent advances in machine learning techniques to generate new insights in support of the ongoing fight against this infectious disease. Even as young software engineering graduates in SJSU, we are trying to do our best in this challenging time. Here are what we do:**

1) **Overview including requirements, architecture, toolkit.**
2) **Select suitable machine learning algorithm on IBM Cloud and establish web UI showing the past COVID-19 situation and predicting the future positive cases based on each state in US.**
3) **Study the reason of COVID-19 outbreak in US and compare several possibilities.**

I. Overview

A. Requirements

We want to build a ML model to predict the trend of the disease, which helps research on the future study of infectious diseases. And also, if accurate, this model can help people to prepare for COVID-19, such as how long they will stay at home in different areas, how many people can be infectious in the future, and which can be dangerous and which are not.

B. Architecture

-Data Set: We collect data from kaggle and github. The data source link is at the end of this file.

-Front End: React.js, Boostrap. Diagram will be simpler and more intuitive to show data relation. As for chart showing, Echart.js is chosen to show our data.

-Back End: Python, Node.js, IBM cloud. We will collect plenty of data associated with COVID-19,then choose most suitable model, train our data in IBM Watson machine learning experiment, and provide an API for prediction.

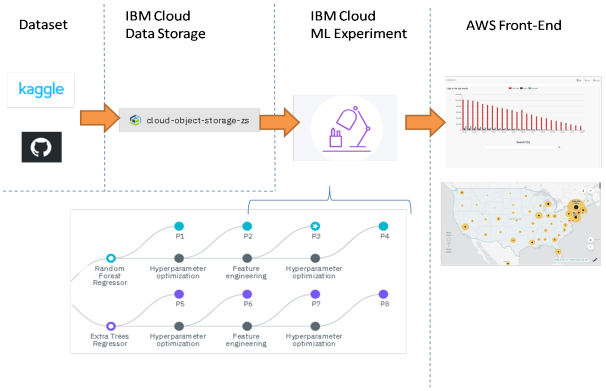


Fig. 1: architecture

We fetch the latest daily data about covid-19 in every state and the nearest one month in USA from Internet, then update them in our front page.

C. Tech Tools

- IBM Cloud Waston Machine Learning
- IBM Cloud Data Storage
- AWS
- ReactJS
- NodeJS
- Python

II. Modeling and UI

A. Select suitable Machine Learning Algorithm on IBM Cloud and upload dataset for computing

We select (datetime, state) as the feature vector in the dataset and new positive population per day as prediction column. IBM Waston Studio provides auto AI development platform. What we need to do is to upload our dataset of COVID-19 to the platform. Then the platform runs and promotes several possible algorithms and calculate the RMSE.

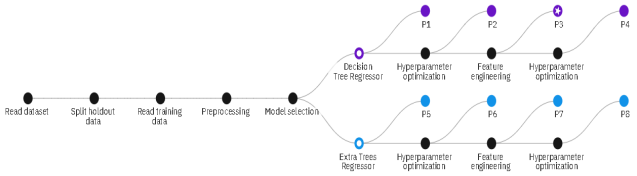


Fig. 2: SelectModel

The algorithm Decision Tree Regressor is selected, for which has the lowest RMSE of 1805. Then we generate a model from it. The Model Evaluation Measures are shown in Table I:

TABLE I: Model Evaluation Measures

	Holdout Score	Cross Validation Score
RMSE	1,292.887	1,805.030
R2	0.997	0.992
MSE	1,671,557.863	3,592,597.296

B. Establish Web UI showing the past COVID-19 situation and predicting the future positive cases based on each state in US

We establish a web site to offer the prediction of each state in US. Users can select the area which they are interested in and select the date they care about. The example is shown in Fig.3.

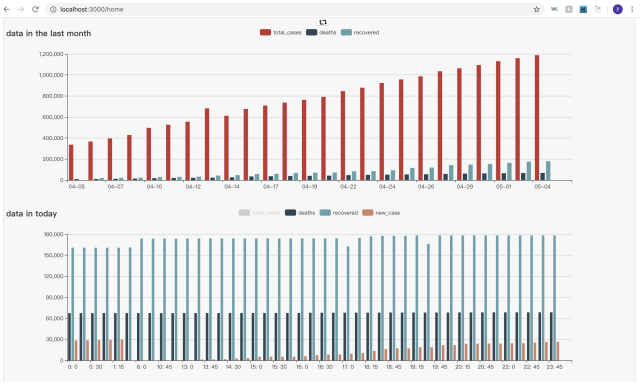


Fig. 3: predict

Also, we provide the current situation of COVID19, as map shown in Fig.4.

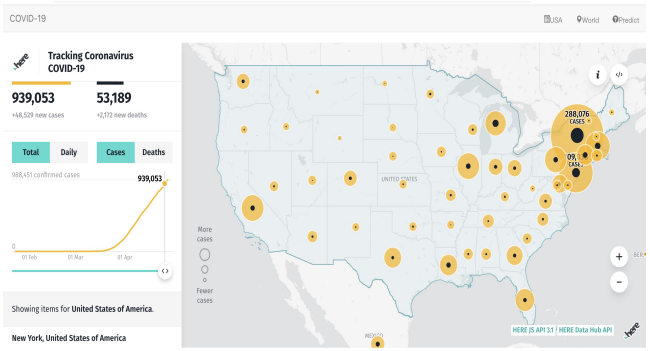


Fig. 4: map

III. Study the reason of COVID-19 outbreak in US and compare several possibilities

A. Wearing Mask Influence on COVID-19

Is wearing a mask helpful for us to prevent from getting the COVID-19?

Does quarantine work for control/reduce the COVID-19 positive increase number for a country?

We all knew that China was the only country which implemented quarantine rule and compulsorily required every citizens to wear mask at the virus exploding period. China had most Covid-19 cases in all over the world, however now China already controlled and weaken Covid-19's spreading speed in China area and set up plan to resume work and resume cities' daily operation.

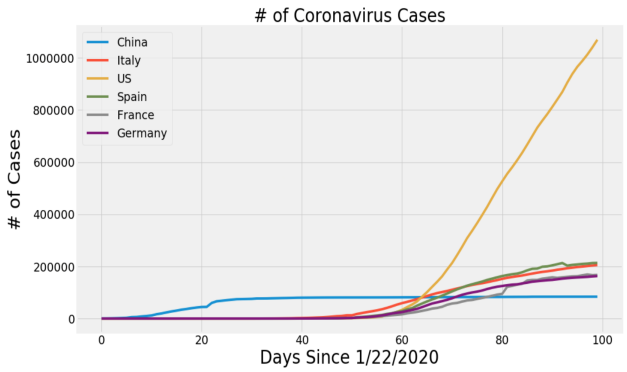


Fig. 5: nations

So, here are two questions: Is wearing mask helpful for us to prevent from getting the COVID-19 Virus? Does the quarantine rule works for control or weaken the Covid-19 positive cases' increase for a country?

We tracked back some news about the quarantine rule and government recommended citizens wear mask as necessary, just as Fig.6.

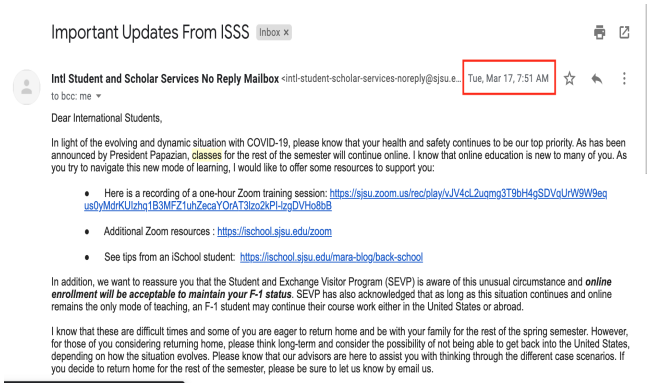


Fig. 6: ImportantUpdates

According to news' date, we could assume that the start date to implement the quarantine rule is Mar. 17th, and the date for CDC and government recommended people wearing facial masks is April 3rd. The dates around these two date (pre or post 2-3 days) also can be considered as the beginning of the implementation of rule suggestion. Then we will check whether data give us clues from these two dates.

Fig.7 and Fig.8 show the data about grocery & retailer mobility corresponding to date. The higher mobility explained the more people go to grocery & retailer compared to last day.

From the data, we can find that, the mobility increased dramatically from Mar.11th to Mar.17th, which also match the panic period in history—people purchased over amounts of grocery to stay at home for long time.

After Mar.17, we can find that the mobility begin to decrease, and the citizens did efficiently cooperate with government's rule.

To make the data more readable, we will fix the fluctuation on the data set. We will calculate the positive increase ratio per day. For example, Day 1, the increase number is 100, Day 2 is 50. Then the increase ratio should be -50%, showed in Fig.9.

According to the data set, we assume that the COVID-19 explode date for the US nation should be on Feb 28th. Before Feb 28th, the number of COVID-19 case still a few. Then we can skip till Feb 28th. This is showed in Fig.10

```
retaildf
```

	location	loc_type	parent_loc	mobility_type	date	mobility_change
0	Afghanistan	parent	world	Retail & recreation	2020-02-16 00:00:00	0.028430
1	Afghanistan	parent	world	Retail & recreation	2020-02-17 00:00:00	0.057307
2	Afghanistan	parent	world	Retail & recreation	2020-02-18 00:00:00	0.025352
3	Afghanistan	parent	world	Retail & recreation	2020-02-19 00:00:00	-0.008219
4	Afghanistan	parent	world	Retail & recreation	2020-02-20 00:00:00	-0.017038
...	...	...	...	...	...	...
942434	Weston County	child	Wyoming	Residential	2020-04-06 00:00:00	-0.387755
942435	Weston County	child	Wyoming	Residential	2020-04-07 00:00:00	-0.367347
942436	Weston County	child	Wyoming	Residential	2020-04-08 00:00:00	-0.360000
942437	Weston County	child	Wyoming	Residential	2020-04-09 00:00:00	-0.312500
942438	Weston County	child	Wyoming	Residential	2020-04-10 00:00:00	-0.431818

Fig. 7: retail0

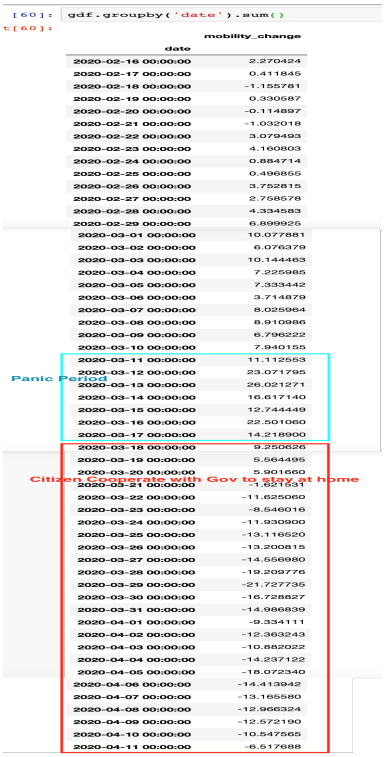


Fig. 8: retail1

The trend of increase rate is not stable, it will be influenced by each day's increase cases dramatically. It would be better for us to count the average to get a period's increase rate. Then we can compare the data between the period before quarantine and after the quarantine, and the period before promoting mask versus after promoting mask.

And we can view that, after the government implements quarantine rule, the increase positive ratio got obvious decrease. From around 45% down to 15%. We can view that the COVID-19

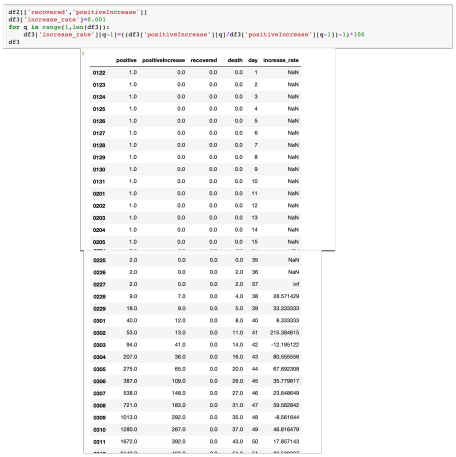


Fig. 9: positiveIncrease

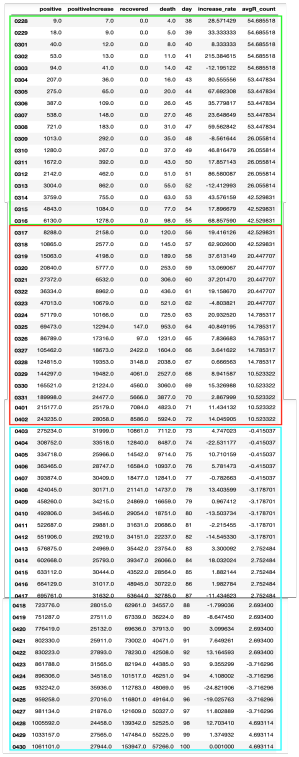


Fig. 10: Feb28th

positive case increase speed gets slow down.

More importantly, after April 3rd, the COVID-19 positive case increase ration from 15% down to below 5%. It is very convincing that quarantine rule/promoting mask have positive effect for a country to fight with COVID-19.

B. US doesn't pass the turning point

The negative increase of the COVID-19 positive case does not mean the US nation already pass the turning point. Actually, the US nation still have not passed the turning point, though some state already passed the turning point.

We have implemented a turning point algorithm. Let us just take the Italy as an example.

	CurrentPositiveCases	NewPositiveCases	day	avg3_count	avg5-count
02-24	221	221	1	130.666667	176.0
02-25	311	93	2	130.666667	176.0
02-26	385	78	3	130.666667	176.0
02-27	588	250	4	242.666667	176.0
02-28	821	238	5	242.666667	176.0
02-29	1049	240	6	242.666667	440.2
03-01	1577	566	7	458.000000	440.2
03-02	1835	342	8	458.000000	440.2
03-03	2263	466	9	458.000000	440.2
03-04	2706	587	10	711.333333	440.2
03-05	3296	769	11	711.333333	1216.6
03-06	3916	778	12	711.333333	1216.6
03-07	5061	1247	13	1512.000000	1216.6
03-08	6387	1492	14	1512.000000	1216.6
03-09	7985	1797	15	1512.000000	1216.6

Fig. 11: Italytp

Because the new increase case get influenced every day, we can count the average increase cases for a period first. We can set the ratio as 3-day a period and 5-day a period. 3-day period could get more in precise date for turning point. However, if the data fluctuates in very huge range, we would recommend to use large scale range.

```
            if ee['avg5-count'][m]==ee['avg5-count'][m-1]:
                Dlist.append(m-1)
                continue
            else:
                Dlist.append(m-1)
                break
        break
ThreeDList = []

for x in range (1,len(Dlist)):
    if ee['avg3_count'][Dlist[x-1]]<=ee['avg3_count'][Dlist[x]]:
        continue
    else:

        ThreeDList.append(Dlist[x-1])
        ThreeDList.append(Dlist[x-1]-1)
        ThreeDList.append(Dlist[x-1]-2)
for  s in range(len(ThreeDList)):
    print(ThreeDList[s])
```

```
32
31
30
```

```
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

Fig. 12: tp

From the turning point algorithm, we know that, the turning point will be appear around the 30-32 row of the dataframe(Fig.12), which is between the date from 3-25 to 3-27.

And we do find it the increase case vector get decrease after the turning point. (Fig.13)

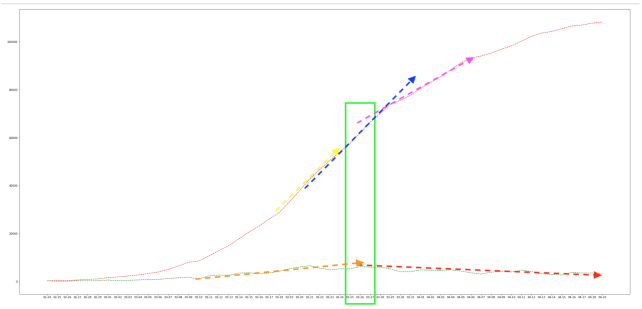
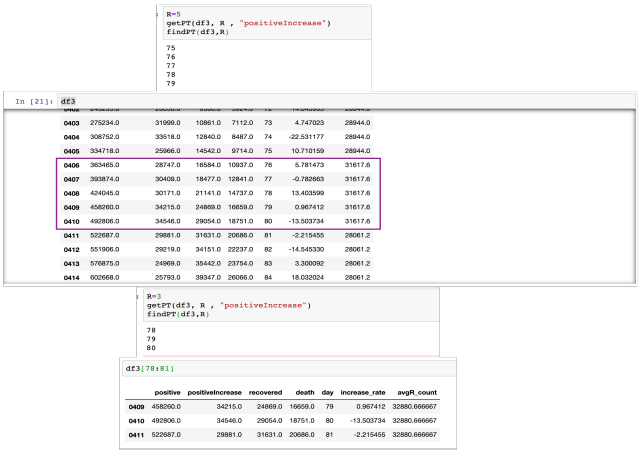


Fig. 13: Line chart

IF we implement this into the US dataset:(Fig.14)



```
R=5
getPT(df3, R , "positiveIncrease")
findPT(df3,R)
```

```
75
76
77
78
79
```

```
In [21]: df3
```

	positive	positiveIncrease	recovered	death	day	increase_rate	avgR_count
0403	275234.0	31999.0	10861.0	7112.0	73	4.747023	28944.0
0404	308752.0	33518.0	12840.0	8487.0	74	-22.531177	28944.0
0405	334718.0	25966.0	14542.0	9714.0	75	10.710159	28944.0
0406	363465.0	28747.0	16584.0	10937.0	76	5.781473	31617.6
0407	393874.0	30409.0	18477.0	12841.0	77	-0.782663	31617.6
0408	424045.0	30171.0	21141.0	14737.0	78	13.403599	31617.6
0409	458260.0	34215.0	24869.0	16659.0	79	0.967412	31617.6
0410	492806.0	34546.0	29054.0	18751.0	80	-13.503734	31617.6
0411	522687.0	29881.0	31631.0	20686.0	81	-2.215455	28061.2
0412	551906.0	29219.0	34151.0	22237.0	82	-14.545330	28061.2
0413	576875.0	24969.0	35442.0	23754.0	83	3.300092	28061.2
0414	602668.0	25793.0	39347.0	26066.0	84	18.032024	28061.2

```
R=3
getPT(df3, R , "positiveIncrease")
findPT(df3,R)
```

```
78
79
80
```

```
df3[78:81]
```

	positive	positiveIncrease	recovered	death	day	increase_rate	avgR_count
0409	458260.0	34215.0	24869.0	16659.0	79	0.967412	32880.666667
0410	492806.0	34546.0	29054.0	18751.0	80	-13.503734	32880.666667
0411	522687.0	29881.0	31631.0	20686.0	81	-2.215455	32880.666667

Fig. 14: US datasheet

We will find it the turning point for the US nation not come yet, but at the period from April 6 to April 10 will have the most increase case per day. After that, later new COVID-19 cases per day have negative increase with very low rate.

Why quarantine rule and wearing mask work, why we still have lot of increase case per day? Previous work, such as the quarantine rule and wearing mask, we just control the increase rate for new COVID-19 cases. If we do not implement those works, right now, we may have not rather 30000 new cases per day, but 50000, 60000 new cases per day.

Why we still don't have turning point yet? Although some states already passed the turning point, the U.S is a very big country, that COVID-19 condition is different from states to states. Until all the states from U.S pass the turning point, then the whole nation will pass the turning point then.

IV. CONCLUSIONS

In conclusion, we just control and reduce the accelerate rate for new cases per day, however there is still a distance for us to ultimately beat the virus for the whole nation.

REFERENCES

[1] https://www.politico.com/interactives/2020/coronavirus-testing-by-state-chart-of-new-cases/

[2] USA data set source: https://www.kaggle.com/sudalairajkumar/covid19-in-usa

[3] https://covid-19.direct/county/CA/Santa%20Clara

[4] https://aqicn.org/data-platform/register/

[5] Fig.5 nations: https://www.kaggle.com/tarunkr/covid-19-case-study-analysis-viz-comparisons

[6] Mobility data set source:https://www.kaggle.com/lanheken0/community-mobility-data

[7] Italy Covid-19 data set source:https://www.kaggle.com/sudalairajkumar/covid-in-italy#covid19_italy_region.csv