UNCOVER COVID-19 Challenge
Which populations assessed should stay home and which should see an HCP?

ABSTRACT

The tasks associated with this dataset were developed and evaluated by global frontline healthcare providers, hospitals, suppliers, and policy makers. They represent key research questions where insights developed by the Kaggle community can be most impactful in the areas of at-risk population evaluation and capacity management.[1]
You can see more information in Kaggle website which attached on the References area.
My juptyer notebook work could be found athttps://github.com/hjjkk/ml_final_project

CONTACT

孔云飞
计算机17-1班
2017011306
Email:2017011306@student.cup.edu.cn

INTRODUCTION

Since 2019-2020,covid-19 has spread all over the world.
In US, the virus has caused huge loss of human, material and financial resources, many people dead in this public health security incident. There is a graph about US epidemic recently underneath.

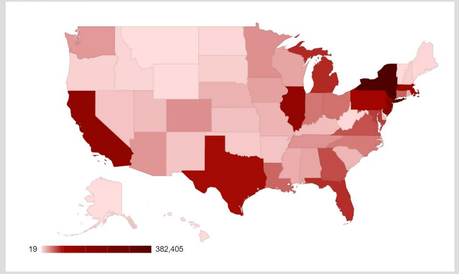

Figure 1. US condition[2]

In this poster, I will analyze data, building a model to predict which populations assessed should stay home and which should see an HCP.
One person should stay home or see an HCP will be judged by if he **is verified** or not.

DATA DESCRIPTION

Filename: coders_against_covid.zip
Number of datasets on Namara: 1
Access streaming data on [Namara](https://app.namara.io/#/search?sources=5e76bb9cbe428d03ee7b1689)
View [source](https://github.com/codersagainstcovidorg/covid19testing-map).
Source description: Crowdsourced map of testing locations across the US.

DATA PROCESSING

First, divide this dataset into two parts:train_data and test_data, 75% of source file into train_data and 25% to be test_data.
Here is description in detail below.

After analysis, some data are useless in modeling.
Such as 0 location_id was hash code example a3b3214a-e128-4c68-ac18-a467482f1ab8

Droping some columns and convert 't' and 'f' into 1 and 0.

```
Data columns (total 9 columns):
 #   Column                                              Non-Null Count  Dtype
---  ------                                              --------------  -----
 0   is_verified                                         3119 non-null   int32
 1   is_hidden                                           3119 non-null   int32
 2   is_location_screening_patients                      3119 non-null   int32
 3   is_location_collecting_specimens                    3119 non-null   int32
 4   lat                                                 3119 non-null   float64
 5   lng                                                 3119 non-null   float64
 6   is_location_accepting_third_party_orders_for_testing  3119 non-null   int32
 7   is_location_only_testing_patients_that_meet_criteria  3119 non-null   int32
 8   is_location_by_appointment_only                     3119 non-null   int32
dtypes: float64(2), int32(7)
```

	is_verified	is_hidden	is_location_screening_patients	is_location_collecting_specimens	lat	lng
3201	0	0	1	0	43.241082	-97.674693
2172	0	0	1	0	45.681274	-111.042090
3934	1	0	1	0	47.712307	-122.339110
3607	0	0	1	0	44.461197	-71.696180
2954	0	0	1	0	35.477990	-97.489370
1999	0	0	1	0	31.130804	-90.138380
189	0	0	1	0	35.206802	-91.733560
1059	1	0	1	1	42.056228	-87.740086
841	0	0	1	0	30.118229	-85.209760
4001	0	0	1	0	39.428731	-78.985996

Figure 2. new table info

ANALYSIS

Due to the propagation characteristics of covid-19, it's easy to find many cases will be confirmed **regionally**,so there're many verified cases in some specific areas.The following pics show some information about the conclusion.

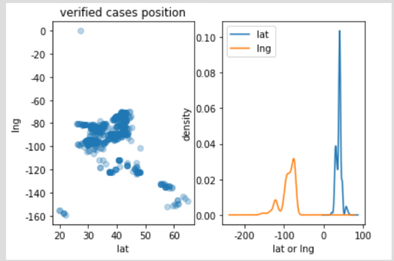

Figure 3. cases in location

Next normalized 'lat' and 'lng' values into 0~1 to avoid too large scale data.
Then show correlations by heatmap.

Figure 4. correlations between different values

According to the figure 4,we can make a preliminary guess that is_location_collecting_specimens,is_location_accepting_third_party_orders_for_testing, is_location_only_testing_patients_that_meet_criteria and is_location_by_appointment_only have high correlation with is_verified.
Next show data relations with is_verified

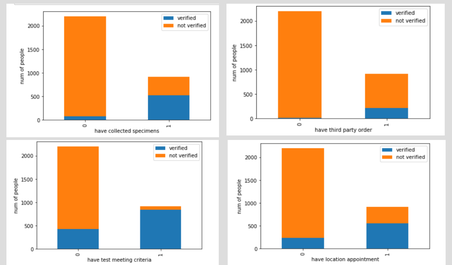

Figure 5

BUILD MODEL

In this step, I use LR(LogisticRegression) to make a model. Optimized hyperparameters with GridSearch.Choosing 7 feature vectors which have close relationship.

```
LogisticRegression(C=1, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=1e-06, verbose=0,
                   warm_start=False)
```

Use 3-cross validation to verify model in train data and draw learning curve as figure 6

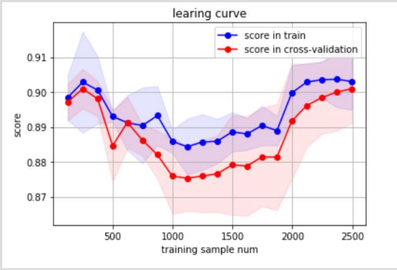

Figure 6

PREDICT

According to the learning curve, this model have good performance in train data, then predict in test data. Using sklearn.metrics to calculate accuracy_score. It has a high accuracy 89.90%

```
from sklearn.metrics import accuracy_score
print(accuracy_score(result_np, predictions))
```

```
0.8990384615384616
```

REFERENCES

1. https://www.kaggle.com/roche-data-science-coalition/uncover/tasks?taskId=674.
2. https://www.guruin.com/guides/covid19