**Are you going to die in COVID-19？**

**Wensong Qiao**

1.Background

As COVID-19 has spread in all the world, people begin to become very worried about it will like black death to create massive infection among people. Although in china, the government make an effort to control the COVID-19 in China, it still spread in other area. In this time, we should research a little bit about the COVID-19 data, and see if there area any insights that can help us to understand more about it.

Although there are lots of CSV files that we can explore, today’s time is limited. I just use COVID19\_line\_list\_data file to explore.

2. EDA

I have done for some of EDA. There are some interesting findings:

1. China has the most infected persons, and then is japan Korean.
2. Wuhan has the most infected persons, and most of highly infected areas are in Asian counties.
3. It seems that male is a little bit easier to be infected
4. Children are not easy to be infected, but seniors are easier to be infected compared with other ages groups.
5. 1085 persons are infected, 160 are recovered, and 62 died.
6. It is just a short time to from exposure to symptom. Usually, people spend 11 days to be diagnostic.

3.Interesting Question

After EDA, I found that some people with specific features are easier to get infected. Therefore, I want to know (1) If we can predict if a person is infected by some features? (2) If we can predict if a infected people will die? (3) If we can predict if a infected people will be recovered?

4. The meaning of the question

For individuals, we can use model to predict the probability to be infected, recovered and die. Therefore, we can have some actions to prevent the bad situation happen.

For the government, we can design a better strategy to control the COVID-19. For example, if there are lots infected people died, we should speed up vaccine development and invest more resource.

5. The future plan

In fact, my questions are about classification problem. This time is limited, I can’t follow a real modeling process to deal with all problems correctly. Even our XGB have an really high accuracy rate, but it not reliable, because it is unbalanced data. For further improvement, I will do more feature engineering to combine different tables to create more features, and use SMOTE to deal with unbalanced data. Finally, use XGB and stacking to create a predictive model.