5741 Project Proposal

Karina Fang, Kornelia Wang, Pengbo Hao

October 3, 2021

Can we predict the possibility of testing positive given the characteristics of a certain patient?

In the year 2020, the world suffered from an unprecedented pandemic COVID-19. Not only has the COVID-19 caused a tremendous amount of death worldwide and significant losses in economies, but this contagious disease is also ever-changing, developing new strains and variants rapidly. Although many methods, including quarantine and testing, have been adopted to prevent the spread of the disease, it can be costly for society if the administration were to perform quarantine and testing on a large scale. Moreover, the number of testing kits is limited and may be distributed unevenly across the country.

To circumvent these constraints, it is vital to answer the question: can we predict the possibility of testing positive given the characteristics of a certain patient. By building a model using the data available, we will be able to prioritize the individuals in need and allocate the testing kits more efficiently. More importantly, the model will provide reliable guidance on when to require an individual to quarantine, and reduce the cost of unnecessary quarantine and excessive testing.

In this project, we want to use the computational method we learned in class to help determine whether there is an approach to identify symptoms to help with early detection of COVID-19 in the United States. To solve this problem, we are planning to build a model which enables us to predict the probability of testing positive for COVID-19 given the person's current symptoms. To fulfill this goal, we choose to use the Aggregated US survey data made available to the public by CMU.

The COVID-19 Symptom Survey designed by Carnegie Mellon University aims at helping researchers better monitor and forecast the spread of the covid. In this survey, participants are invited to self-report COVID-19-related symptoms such as cough, sore throat and difficulty breathing, etc. Other information including gender, age and chronic diseases are also recorded to give the researcher a better sense of the background of the participants.

With all aggregated datasets provided by CMU, we are planning to use the overall-state-smoothed data to build our model. The dataset has 16290 entries in total, with 105 independent variables involving basic information and symptoms of the participants. All variables in the dataset are clearly listed and none of them need special knowledge to be understood. Among all these independent variables, only three of them (state code, gender, age bucket) are categorical variables and the rest are all continuous variables. These continuous variables represent the proportion of participants having a certain symptom among all participants in each entry. The overall quality of this dataset is great, with only a few entries of missing data.

In conclusion, the dataset would enable us to develop new perspectives to answer our question. The relationship between testing results and significant symptom signals can be discovered by interpreting the coefficients, and the model can be improved by picking different model classes, which would eventually lead us to find an effective model to answer the question.

dataset: https://cmu.app.box.com/s/ymnmu3i125go4aue0qxosi3rbcae20bj