ORIE5741 Midterm Report

Karina Fang, Kornelia Wang, Pengbo Hao

November 1, 2021

1 Introduction

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.

2 Data

2.1 Data Descirption

For this project, we use the Aggregated US survey data made available to the public by CMU. This dataset is collected by Facebook, and sampled users receive the invitation at the top of their News Feed and the surveys are conducted off the Facebook app. Every day, a new sample of Facebook users over 18 years old are invited to consent and participate in the survey and self-report their COVID-19-related symptoms. Using this survey data, we estimate the percentage of people who have a COVID-like illness, in a given location, on a given day. All numerical data points are adjusted to be the 7 day trailing average (today, and the 6 previous days).

The data originally has 162691 rows and 105 columns, containing 8 dimension columns and 50 signals. Among all the columns, the testing result columns, 35, 36, 37, 38 and 39, are our labels; the rest of the columns are our features.

2.2 Data Cleaning

We generate a data matrix, use column 35 as a target vector, and create a train/test split so that we can empirically test for overfitting. To make the 80 / 20 train test split, we shuffle the data and select the first 80% as the train data, with 20% held out for validation.

2.3 Feature Selection

For feature selection, first we consider numerical variables that might be useful for predicting the results: Estimated percentage of people reporting that they have experienced COVID-19-related symptoms. These symptoms include fever, cough, short of breath, difficulty breathing, tiredness or exhaustion, nasal congestion, a runny nose, muscle or joint aches, a sore throat, persistent pain or pressure in their chest, nausea or vomiting, diarrhea, loss of smell or taste, etc.

We also consider categorical variables: gender and age_bucket, as these variables are also informative. Since possible values of gender and age_bucket also contain "overall" (aggregation of all genders), we delete these rows to avoid duplication. Then we transform the data into one-hot vectors and concatenate those features with numerical data we had previously.

We only took the examples for which the result of the estimated percentage of people reporting that they have tested positive for COVID-19 is known into consideration. Since this survey item is only

presented to respondents who report they are currently experiencing symptoms, most responses are blank.

After data cleaning, we have 45381 rows x 100 columns in our data matrix. No data is missing in the feature matrix.

3 Data Analysis

3.1 Data Visualization

Data visualization is an important step to identify the important features that are useful in building the model and making predictions. We first visualized the correlation between different features to gain a better understanding of how different features might influence the probability of testing positive of a patient. Then, by plotting histograms of selected important features, we visualized the underlying distribution of selected features, which gave rise to the construction of the model.

3.1.1 Correlation Visualization

As we can observe in Figure 1.1, the dependent variable "smoothed_pct_tested_and_positive" is highly correlated to feature "smoothed_pct_self_fever" (7-day trailing average of the estimated percentage of people reporting that they have experienced a fever in the past 24 hours), "smoothed_pct_self_anosmia_ageusia"(7-day trailing average of the estimated percentage of people reporting that they have experienced the loss of smell or taste in the past 24 hours). Our dependent variable is also observed to be weakly correlated to feature "smoothed_pct_self_difficulty_breathing" (7-day trailing average of the estimated percentage of people reporting that they have experienced difficulty breathing in the past 24 hours), and "smoothed_pct_self_nausea_vomiting" (7-day trailing average of the estimated percentage of people reporting that they have experienced nausea or vomiting in the past 24 hours). All four features are aligned with the common belief of the symptoms of COVID-19. To better understand the data, we further plotted the histogram of each feature.

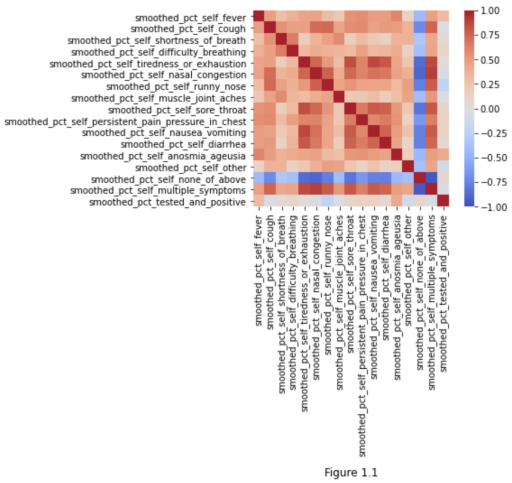
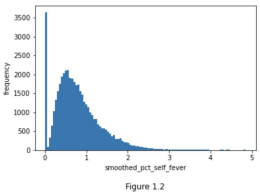


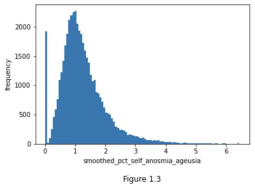
Figure 1: correlation visualization

3.1.2 Histogram Visualization

Before plotting the histogram, we first provide the summary data for the four features of interest in Table 1.1.

Then, we plotted the histogram for each feature of interest and observed the underlying distribution of the data. As we can observe in Figure 1.2, while there are a significant amount of people reporting no sign of fever in the past 24 hours, the remaining percentage of people reporting a fever is approximately log normal distributed with mean around 0.8%. Similarly, from Figure 1.3, there are a moderately large amount of people reporting no experience of loss of smell or taste in the past 24 hours, the rest of the data suggesting an approximate log normal distribution with mean around 1.2%. However, for the features that are weakly correlated to the dependent variable, as suggested in Figure 1.4 and Figure 1.5, the distribution of the data remains unclear. In order to have a better understanding of the data, we conducted preliminary analyses on the data utilizing linear regression.





	smoothed_pct_self_fever	smoothed_pct_self_anosmia_ageusia	smoothed_pct_self_difficulty_breathing	smoothed_pct_self_nausea_vomiting
count	45381.000000	45381.000000	45381.000000	45381.000000
mean	0.784458	1.239205	2.251551	2.642545
std	0.547343	0.737408	0.871985	1.767999
min	0.000000	-0.000000	-0.000000	0.000000
25%	0.412200	0.772200	1.686700	1.292600
50%	0.689300	1.108800	2.175200	2.168300
75%	1.058600	1.582000	2.741900	3.689400
max	4.850100	6.451600	9.579800	10.728700

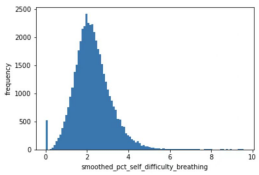


Figure 1.4

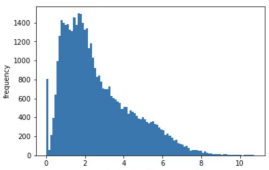


Figure 1.5

3.2 Preliminary Analysis

For the preliminary analysis, we fit a linear model on the data with the features selected from the data clean up step as well as an offset term. After fitting the linear model, we also computed the MSE for both training and testing data. We then run the OLS function and bypass any SVD convergence errors by refitting the model.

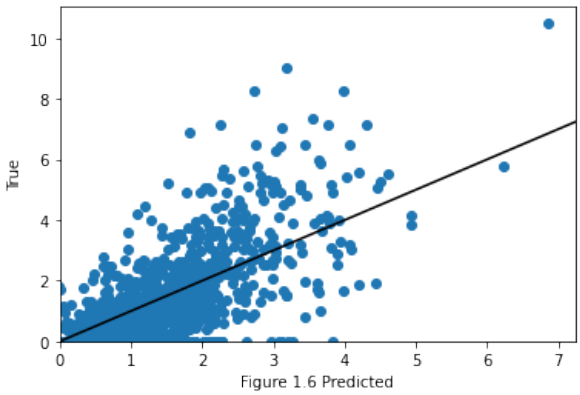


Figure 1.6 Predicted

Figure 1.6 shows our linear prediction, with a training MSE of 1.4434 and a testing MSE of 1.4465. The training MSE and testing MSE are similar with each other and they are low based on our original dataset. Therefore, the model we currently fit is neither overfitting nor underfitting.

4 Future Steps

In our current analysis, we have made some preliminary models with linear regression to predict the probability of testing positive for covid-19. We tested the effectiveness of our model by splitting our data into training and testing sets and calculated the mean squared errors for evaluation.

In the future, we plan to continue this strategy and incorporate k-fold cross validation and apply methods like control burn and bagging to add the complexity of our model. We will also experiment with regularization and use different loss funcitons such as hinge loss and logistic loss to update our prediction model. At the same time, we plan to try other methods to have a better evaluation on the model.