A close up of a sign

Description automatically generated

PPE Guide: Covid-19 in the USA

Professor: Dr. Jiming Wu

Chenwei Cao

July 22, 2020

Table of Contents

[Introduction 3](#_Toc46011846)

[Tools and Datasets 4](#_Toc46011847)

[Data Overview 5](#_Toc46011848)

[Data Pre-Processing 7](#_Toc46011849)

[Missing data processing 7](#_Toc46011850)

[Outlier Processing 8](#_Toc46011851)

[Data Visualization 8](#_Toc46011852)

[Data Analysis 12](#_Toc46011853)

[Model Selection 12](#_Toc46011854)

[Logical Regression 12](#_Toc46011855)

[Long short-term memory/Recurrent Neural Network 13](#_Toc46011856)

[Autoregressive Integrated Moving Average Model 14](#_Toc46011857)

[Model Evaluation 15](#_Toc46011858)

[Why do we need a model evaluation? 15](#_Toc46011859)

[Evaluate the Models 15](#_Toc46011860)

[Conclusion 19](#_Toc46011861)

[References 21](#_Toc46011862)

# Introduction

Novel Coronavirus has affected every aspect of our lives, what is Novel Coronavirus? Coronaviruses are a large family of viruses that may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19. People can catch COVID-19 from others who have the virus. The disease spreads primarily from person to person through small droplets from the nose or mouth, which are expelled when a person with COVID-19 coughs, sneezes or speaks. These droplets are relatively heavy, do not travel far and quickly sink to the ground. People can catch COVID-19 if they breathe in these droplets from a person infected with the virus [1]. As we knew, COVID-19 is highly contagious, and now novel coronavirus variants have emerged. What are the symptoms of COVID-19?

A screenshot of a cell phone

Description automatically generatedA close up of a piece of paper

Description automatically generated

Figure1

Figure 1 shows the symptoms of the COVID-19 table I created. Base on the clinical symptoms of COVID-19 from WHO, we can see that 87.9% of confirmed cases have a fever, 67.7% of confirmed cases have a dry cough, 38.1% of confirmed cases have fatigue, and 33.4% of confirmed cases have sputum production, and so on.

A close up of text on a white background

Description automatically generated

Figure2

To make the symptoms more intuitive, I put all the symptoms of COVID-19 together to make a pie chart. From figure2, we can see that fever, dry cough, and fatigue account for a large proportion. As these graphs show in figure 1 and figure 2, the most common symptoms of COVID-19 are fever, dry cough, and fatigue.

In this project, I have collected, organized, and analyzed some data to give people a more detailed and in-depth understanding of the United States' COVID-19 situation, and predict future COVID-19 trends about confirmed cases.

# Tools and Datasets

The project uses Python and Jupyter notebook as data analysis tools. It also uses Keras API in this project to help build models. Keras is a deep learning API written in Python, running on top of TensorFlow's machine learning platform. It developed with a focus on enabling fast experimentation. Going from idea to result as fast as possible is key to doing proper research [2]. Additionally, the datasets I used are time\_series\_covid19\_confirmed\_global.csv , time\_series\_covid19\_deaths\_global.csv and csse\_covid\_19\_daily\_reports of the USA from COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University, which will update the data day by day.

# Data Overview

A screenshot of a social media post

Description automatically generated

Figure 3

Here is a dataset overview of 'time\_series\_covid19\_confirmed\_global.csv’. As figure 3 shows, we have 266 rows of records, 183 columns; the records show the number of confirmed cases from January 22, 2020 until to July 22, 2020 in all over the world.

A screenshot of a computer

Description automatically generated

Figure 4

As figure 4 shows, in data ‘time\_series\_covid19\_deaths\_global.csv’, we have 266 rows of records, 187 columns. the records show the number of deaths cases from January 22, 2020 until to July 22, 2020 in all over the world.

A close up of a piece of paper

Description automatically generatedA close up of text on a white background

Description automatically generated

Figure 5

Here is a dataset overview of ‘csse\_covid\_19\_daily\_reports of USA’. As figure 5 shows, we have 3234 rows and 14 columns; each column's records are not null, and there are 6 variables in float format, 5 variables are in object format and 3 variables in an integer format.

A screenshot of a cell phone

Description automatically generated

Figure 6

For each variable in data 'csse\_covid\_19\_daily\_reports of USA', we have 3234 counts of confirmed cases, deaths cases, and recovered cases for each county. According to the dataset, we can get the mean of confirmed cases, and deaths cases are 1147.62 and 43.32; standard deviation for each are 6310.61 and 451.50; minimum value for confirmed cases is 0.00, maximum is 221121.00, 25% is 35.00, 50% is 127.50, and 75% is 498.00. minimum value for deaths cases is 0.00, maximum is 23388.00, 25% is 0.00, 50% is 2.00, and 75% is 11.00. From the dataset description, I learned that some columns we do not need, such as 'FIPS', 'Lat', and 'Long'. I have to do data pre-processing to clean up the data, which can help build models more accurately.

# Data Pre-Processing

## Missing data processing

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 7

As figure 7 shows, there is no missing data in data ‘time\_series\_covid19\_confirmed\_global.csv’ and ‘time\_series\_covid19\_deaths\_global.csv’. In data 'csse\_covid\_19\_daily\_reports of USA', we got 11 missing records for FIPS, 6 missing records for the county, 64 missing records for Lat, 64 missing records for Long, 1 missing records for Active, 64 missing records for Incidence rate and 45 missing records for case fatality ratio. Moreover, there is no missing data in the dataset columns Province\_State, Country\_Region, Last\_update, Confirmed, Deaths, Recovered, and Combined\_key. What we need columns for this project are Province\_State, Country\_Region, Last\_update, Confirmed, Deaths. The best way to deal with the missing data is to delete them in this project, which will not affect the models accurately.

## Outlier Processing

A screenshot of a cell phone

Description automatically generated

Figure 8

As figure 8 shows above, In data ‘csse\_covid\_19\_daily\_reports of USA’, we have 51 outliers in column ‘FIPS’, 103 outliers in column ‘Lat’, 26 outliers in column ‘Long’, 24 outliers in column ‘Confirmed’, 14 outliers in ‘Deaths’, 1 outlier in ‘Recovered’, 6 outlier in ‘Active’, 55 outliers in ‘Incidence\_Rate’, and 47 outliers in ‘Case-Fatality\_Ratio’.

Suppose there are outliers in the dataset. There are four ways to deal with outliers:

1. Drop the outlier record to keep them from skewing the analysis
2. Cap the data
3. Assign a new value, in other words, and we can use the mean of a variable or regression model to predict the missing value.
4. Try a transformation, try creating a percentile version of the original field, and work with that new field instead.

In the dataset, the primary columns I used are 'Confirmed' and 'Deaths', so I will ignore other columns' outliers. Since the dataset records are for different states and counties with different populations, we can ignore the outliers for 'Confirmed' and 'Deaths'.

# Data Visualization

After data pre-processing, Next, this project will use data visualization to more visually show the current situation of COVID-19 in the United States.

A close up of a map

Description automatically generated A screenshot of a social media post

Description automatically generated

Figure 9

Base on the trend chart, we can get a COVID-19 confirmed cases trend chart, and a death cases trend chart as figure 9 shows above. Both of them are increasing with the time goes on. Base on the table, we can see that the total confirmed cases are 3970085, and the death cases are 143190 on July 22, 2020 in the USA.

A screenshot of a cell phone

Description automatically generated

Figure 10

The bar chart shows the top ten states with most affected by COVID-19 by July 22, 2020, and we can see that the top ten states with the large number of confirmed cases of COVID-19 include New York, California, Florida, Texas, New Jersey, Illinois, Georgia, Arizona, Massachusetts, and Pennsylvania.

A screenshot of a cell phone

Description automatically generated

Figure 11

The bar chart shows the top ten states with most deaths of COVID-19, we can see that these states with the large number of death cases include New York, New Jersey, Massachusetts, California, Illinois, Pennsylvania, Michigan, Florida, Texas, and Connecticut.

Here is a map of COVID-19 in the United States:

A close up of a map

Description automatically generated

Figure 12

As figure 12 shows, we can see that Covid-19 primarily distributed in the eastern and west coasts of the United States.

To give a better picture of the spread of the epidemic in the United States, this project trying to use a heat map. Heat maps are an excellent tool for visualizing complex statistical data. Here is a heat map of COVID-19 in the United States:

A close up of a map

Description automatically generated

Figure 13

As figure 13 shows, Deep red states with high numbers of confirmed cases. There are six levels of colors from lite red to deep red. Each level from value 0 to value 5.61 represents 0 confirmed cases, 6699 confirmed cases, 36034 confirmed cases, 80337 confirmed cases, 365971 confirmed cases, and 421286 confirmed cases.

# Data Analysis

Before selecting the model to prediction, I used the correlation matrix to show correlation coefficients between variables in the data 'csse\_covid\_19\_daily\_reports of USA’. The correlation matrix used to summarize data as an input into a more advanced analysis and diagnostic for advanced analyses.

A screenshot of a cell phone

Description automatically generated

Figure 14

Since actual green means a high correlation, lite green means a low correlation. From the correlation matrix (Figure 14), we can learn that the variable 'Confirmed' has a high correlation with variable 'Deaths', which means that ‘Deaths’ will change as ‘Confirmed’ changes.

A close up of text on a white background

Description automatically generated

Figure 15

As the correlation plots on figure 15 and the trend chart on figure 9 showed, with the number of confirmed cases increases, so does the number of deaths.

# Model Selection

Next, this project will select different models to predict future COVID-19 confirmed cases based on data from the United States:

## Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous. Like all regression analyses, logistic regression is a predictive analysis. Logistic regression describes data and explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables [3]. I chose logistic regression as a predictive model to analyze and predicted the confirmed cases in the USA.

A close up of a map

Description automatically generated

Figure 16

As figure 16 shows, we can get the prediction of confirmed cases for next seven days in the USA by logistic regression model. The forecast confirmed cases increase with time goes by. Here are results of next seven days’ prediction: there will be 4.184 million confirmed cases on July 23 2020, 4.252 million confirmed cases on July 24 2020, 4.321 million confirmed cases on July 25 2020, 4.392 million confirmed cases on July 26 2020, 4.463 million confirmed cases on July 27 2020, 4.535 million confirmed cases on July 28 2020, and 4.609 million confirmed cases on July 29 2020.

## Long Short-term Memory/Recurrent Neural Network

Next, I select Long short-term memory/recurrent neural networks as a predictive model for predicting confirmed cases in the USA. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in deep learning. LSTM networks are well-suited to classifying, processing and making predictions based on time series data. There can be lags of unknown duration between essential events in a time series [4]. Since all the datasets are time series data in this project, LSTM/RNN will be a good choice to handle them.

A screenshot of a social media post

Description automatically generated

Figure 17

Figure 17 shows prediction results for the next seven days confirmed cases in the USA by LSTM/RNN model. Base on the LSTM/RNN model’s predictions, We can see that there will be 3.996 million confirmed cases on July 23 2020, 4.024 million confirmed cases on July 24 2020, 4.049 million confirmed cases on July 25 2020, 4.066 million confirmed cases on July 26 2020, 4.078 million confirmed cases on July 27 2020, 4.085 million confirmed cases on July 28 2020, and 4.092 million confirmed cases on July 29 2020.

# Model Evaluation

## Why do we need a model evaluation?

Model evaluation is an integral part of the model development process. It helps to find the best model representing our data and how well the chosen model will work in the future.

## Evaluate the Models

For the actual model training, I build Logistic Regression and LSTM/RNN models by Keras API in Python.

First of all, I will evaluate the Logistic Regression model.

As figure 18 shows that I divided my data into training set and testing set. Training set is the set of data used to train the model. For each epoch, our model will be trained over and over again on it. The purpose of training the model is that we can deploy our model, and have it accurately predicted on new data that it’s never seen before. Testing set is the set of data that is used to test the model after the model has already been trained and validated after using our training set and validation set. The major difference between the testing set and the two other sets is that the test set should not be labeled. The training set and validation set have to be labeled so that we can see the metrics given during training, like the loss and the accuracy from each epoch.

A screenshot of a cell phone

Description automatically generated

Figure 18

We can learn that there are 6721 params, 6721 trainable params, and 0 non-trainable params for the logistic regression model.

A screenshot of a cell phone

Description automatically generated

Figure 19

As figure 19 shows above, I divided the data into 70% of data as training set and 30% of data as testing set. Training set is the set of data used to train the model. For each epoch, our model will be trained over and over again on it. The purpose of training the model is that we can deploy our model, and have it accurately predicted on new data that it’s never seen before. Testing set is the set of data that is used to test the model after the model has already been trained and validated after using our training set. The major difference between the testing set and the two other sets is that the test set should not be labeled. The training set has to be labeled so that we can see the metrics given during training, like the loss and the accuracy from each epoch.

A screenshot of a social media post

Description automatically generated

Figure 20

Here I used build in function called fit () to train the model. The loss represents to mean squared error, which is a scalar value that we attempt to minimize during our training of the model. We can use loss value in Keras on regression problems. The lower the loss, the closer our predictions are to the actual values. The accuracy,0.0109, is used to compute the frequency with which y prediction matches y actual which are used to evaluate classification models. For a regression model, what we can use to evaluate our model is loss value. Base on the model output history, we can see that the loss is decreasing with the epoch increasing, which means there are so we can say that the regression model results are more and more close to actual values.

Secondly, we will evaluate the LSTM/RNN model.

A screenshot of a cell phone

Description automatically generated

Figure 21

Here is the summary for the LSTM/RNN model, we can see that the total params are 50851, the trainable params are 50851, and the non-trainable params is 0.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

Figure 22

For the LSTM/RNN model, the loss represents the mean squared error. Base on the chart about LSTM/RNN model’s loss during train the model, we can see that with the epochs increasing, the loss is decreasing. which means the LSTM/RNN model is more and more accurate.

Base on the two models results, with same input data and epoch number for the two different models, the mean square error is lower in LSTM/RNN model than the mean square error in Logistic regression, we can say that LSTM/RNN model’s predictions are more accuracy than the logistic regression model’s predictions.

# Conclusion

This project introduced what COVID-19 is and what symptoms of COVID-19 are and used dataset about COVID-19 of the USA from Data Repository by the Center for Systems Science and Engineering. It also did data pre-processing base on the datasets to deal with missing data and outlier data. Additionally, it did data visualization to make the data more intuitive. People can more easily understand the distribution and situation of COVID-19 in the United States. For predictive part, there are logistic regression and LSTM/RNN models to predict the USA's confirmed cases in the project. Additionally, it also contains model evaluation process, which can help people learn which model is more accuracy.

# References

[1] WHO, “Q&A on coronaviruses (COVID-19)” WHO. [Website]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses, Accessed on: Jun. 15, 2020.

[2] Keras, “Deep learning for humans” Keras. [Website]. Available: https://keras.io/, Accessed on: Jun. 15, 2020.

[3] StatisticsSolutions, “What is Logistic Regression?” StatisticsSolutions. [Website]. Available: https://www.statisticssolutions.com/what-is-logistic-regression/, Accessed on: Jun 30, 2020.

[4] Wikipedia, “Long short-term memory” Wikipedia. [Website]. Available: https://en.wikipedia.org/wiki/Long\_short-term\_memory, Accessed on: July 15, 2020.

# 