BIOS 823 Final Project: Prediction of COVID-19 cases using machine learning approach

Team UUS: Yiyang Zhang, Candice Li, Jiajie Shen

Abstract

Ever since the first case was diagnosed at the end of 2019, the Covid-19 coronavirus, or SARS-CoV-2, has severely affected human health and social economic stability on a global scale. Covid-19 is an infectious disease that spreads rapidly in humans through the air. It is necessary to analyse the treatment and prevention measures by tracking the trend of new cases and total deaths cases. Machine learning techniques are now being widely used and are playing an more important role in predictive science among medical fields than ever. We have seen a rising usage of supervised machine learning models with their associated algorithms to analyze large biomedical and demographic datasets for either regression or classification. They are contributing to modern biomedical research by training various models to predict deaths and confirmed cases of new diseases. In this project, a dataset of the Covid-19 cases in the United States is being collected, pre-processed, and the corresponding models are being trained by supervised machine learning algorithms to predict the trend of the total Covid-19 confirmed cases in the near future.

1. Introduction

1.1. Data collection

We have taken our study dataset from the website of the Center of Disease Control (CDC) since open access of related data has been given by CDC regarding Covid-19. The study dataset we chose is a time series dataset which focuses on the cases and deaths counts by state in the United States of America.

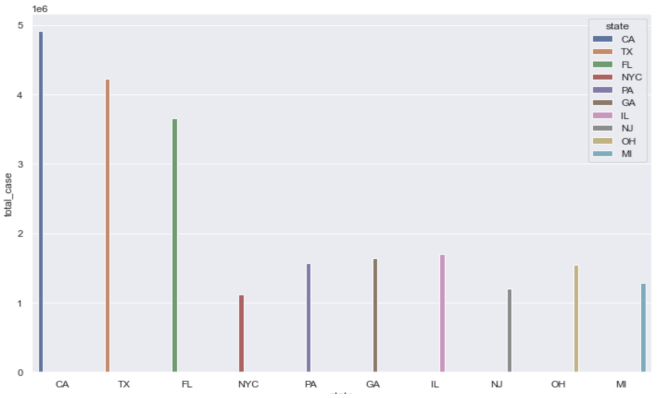
1.2. Preprocessing

Since this dataset has been well collected by medical professionals, there are little to no missing values. Thus we did not perform any imputation measures. We have transformed this time series dataset into a supervised learning dataset by setting and changing the shapes and attributes of variables in our dataset. Since we already have a mass release of various Covid-19 vaccines, we want to discover the trend of Covid-19 spread with vaccinations taken into account. Thus we set the starting point of our dataset as December 14th, 2020, which is the date of the very first covid vaccine got approved in the United States of America.

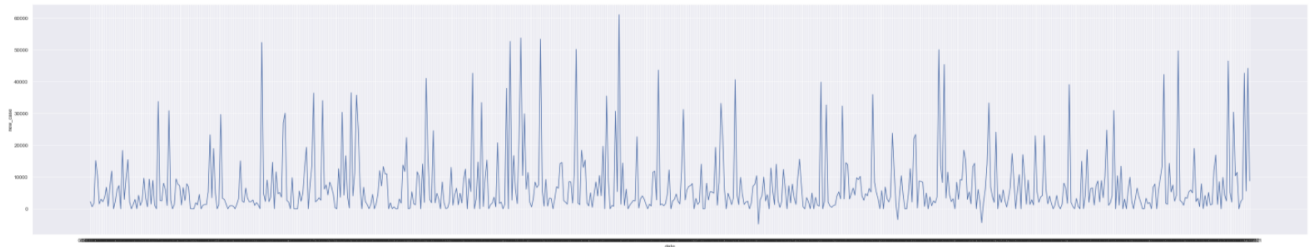
1.3. Feature extraction

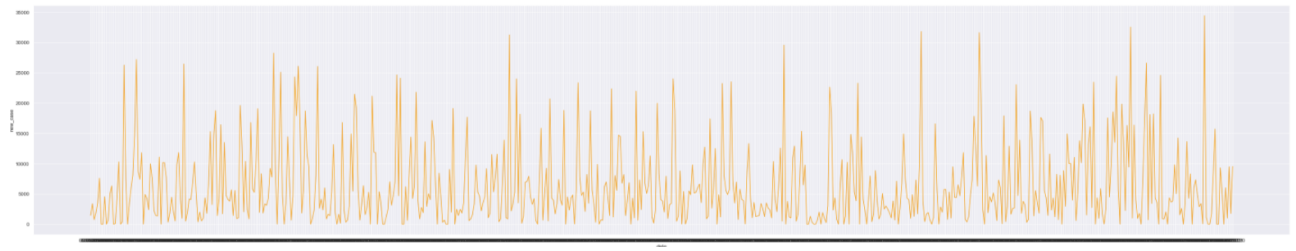
Various features have been extracted from the preprocessed dataset by using NumPy and Pandas libraries. Relevant features are identified as ***date, state, total cases, new cases, total deaths*** and ***new deaths***. We eventually decided that ***date, state,*** and ***total cases*** are our final features.

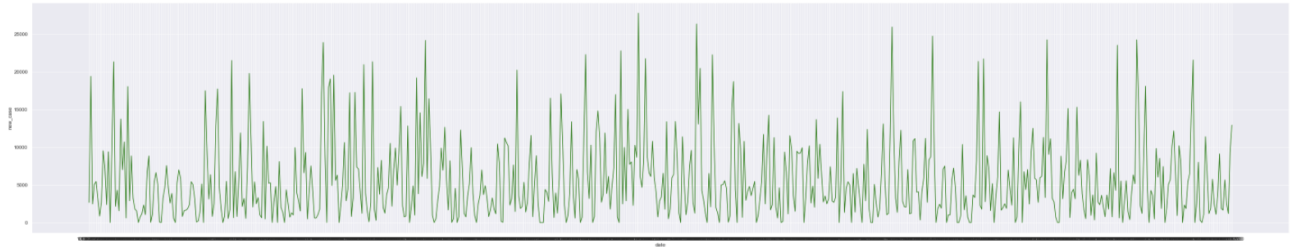
1.4. Exploratory Data Analysis



Since we want to study the trend of Covid-19 in a more detailed fashion, we selected the top three states with the most deaths. These three states are California(CA), Texas(TX) and Florida(FL). These three states have significantly more deaths cases than other states from our observation. Then we look at the overall trend of Covid-19 cases in the three states. From the plots, we can not conclude anything.







2.Prediction methodologies

Predicting Covid-19 cases

To forecast the number of new cases, we employ ridge regression and supervised machine learning algorithms like random forest regressor to train the model by providing the necessary features as the training input. Then we try to fit a multi-layer perceptron model to our dataset to predict for our home state of North Carolina.

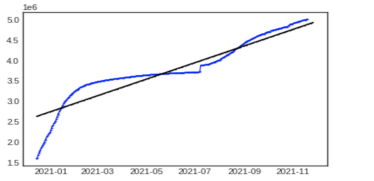
3.Corresponding results and visualizations

3.1. Ridge Regression

The Ridge Regression model for predicting future COVID-19 total cases in CA, TX and FL performs with an average RMSE of 132687.106, 360739.009 and 362609.554. The corresponding *r*2 and MAE values in each state's cases are CA(0.01583,132390.16), TX(-6.60539, 358889.28) and FL(-31.07373, 355831.96). Based on our prediction, in the time period from 11.19.2021 to 11.26.2021(7 days after the end date of our dataset), the mean of total Covid-19 cases in each state is: 4913282.59 (CA), 3984442.84 (TX), and 3489527.21(FL). Plots down below depict the details and accuracy of the models' predictions.

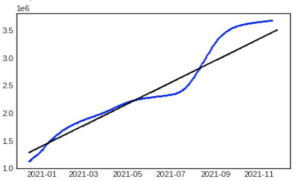
```
# california
forecast_r_ca, new_date_time_index_ca, model_score_
```

```
RMSE for Ridge Regression Regression:  [132687.1061
Mean Absolute Error:  132390.1686492473
R2-Squared: 0.015833285852311363
```



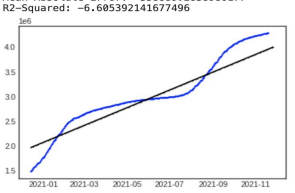
```
# Florida
forecast_r_fl, new_date_time_index_fl, model_score_r_fl = R
```

```
RMSE for Ridge Regression Regression:  [362609.5535802476]
Mean Absolute Error:  355831.9658269108
R2-Squared: -31.073739984834212
```



```
# Texas
forecast_r_tx, new_date_time_index_tx, model_score_r_tx = I
```

```
RMSE for Ridge Regression Regression:  [360739.008769759]
Mean Absolute Error:  358889.2838986177
R2-Squared: -6.605392141677496
```

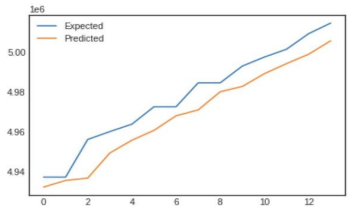


3.2. Random Forest Regression

The random forest regressor model for predicting future COVID-19 total cases in CA, TX and FL performs with an average MAE of 8870.105, 5160.919 and 2424.101. The corresponding *r*2 values in each state's cases are CA(0.82806), TX(0.83538) and FL(0.84961). Based on our prediction, we only got a flat curve for future cases in the next 7 days period, meaning that there were some issues with our model. Plots down below depict the details and accuracy of the models' predictions.

```
# Random forest-California
```

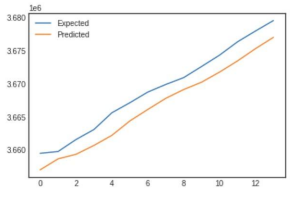
```
>expected=4937188.0, predicted=4932200.4
>expected=4937188.0, predicted=4935532.8
>expected=4956081.0, predicted=4936692.6
>expected=4960026.0, predicted=4949281.5
>expected=4963742.0, predicted=4955683.5
>expected=4972469.0, predicted=4960634.1
>expected=4972469.0, predicted=4967925.9
>expected=4984465.0, predicted=4970874.9
>expected=4984465.0, predicted=4979996.6
>expected=4992875.0, predicted=4982649.1
>expected=4997390.0, predicted=4989051.0
>expected=5001302.0, predicted=4994089.6
>expected=5009101.0, predicted=4998785.6
>expected=5014428.0, predicted=5005544.6
[4932200.403, 4935532.799, 4936692.647, 494928
MAE: 8874.764
```




```
eval_reg(y, yhat)
```

```
Mean Absolute Error:  8870.10578571433
R2-Squared: 0.8280642147192246
```

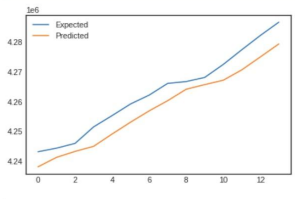
```
# Random forest-Florida
```




```
eval_reg(y, yhat)
```

```
Mean Absolute Error:  2424.1014285714564
R2-Squared: 0.8496109618785681
```

```
#Random forest-Texas
```

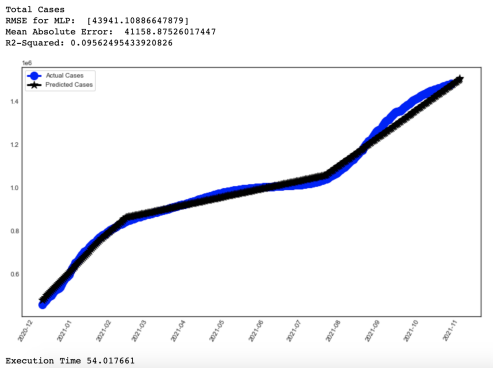


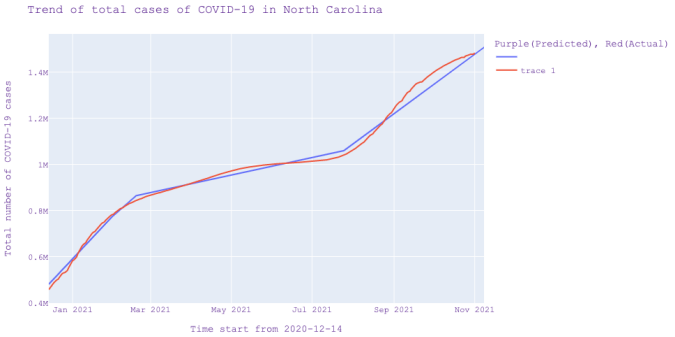

```
eval_reg(y, yhat)
```

```
Mean Absolute Error:  5160.919142857146
R2-Squared: 0.8353895233136869
```

3.3. Multi-layer Perceptron

The multi-layer perceptron model for predicting future COVID-19 total deaths in NC performs with an RMSE of 43941.108, an MAE of 41158.875 and an *r*2 of 0.0956. Based on our prediction, we only got a curve that is going upward for future cases in the next 7 days period, meaning that there is an increase in the trend of new Vocid-19 cases in North Carolina. The average Covid-19 cases in the next 7 day period of our prediction is 1498324 per day.





Click the link below for an interactive line plot of the trend & prediction:
https://share.streamlit.io/sl679/823_final_project/main/MLP_FinalProject_for_streamlit.py

4. Conclusions and future works

The ridge regression approach gives us some low $r2$ scores, which means that the models do not fit very well. The random forest regressor, on the other hand, provided us with some good $r2$ scores, but it failed to perform functionally as only a flat curve was shown for future cases in the next 7 days period. This means that there were some issues with our random forest model. We think that it is due to the fact that the largest value of our threshold is not about to split the values at the last tree node, it always ends up in the same node which results in the same prediction value. For the multi-layer perceptron model, since we only have limited computational power, we failed to do a thorough gird search. We tried a couple combinations of hyperparameters and chose the ones that performed the best. The error metrics were not very optimal either compared to the other models we did. For the random forest model, we plan on trying to add more variables like new deaths confirmed so that we have more features to form a random forest model that is more efficient. We also plan to try a more inclusive and thorough grid search to provide a more optimal result on our prediction. In addition, some new methods/models like ARIMA could be fitted to improve the overall performance of our analysis.

5. Individual contributions and reflections

Candice Li:
In this project, I built a Ridge regression model and a Random forest for the total cases of confirmed COVID 19 in California, Florida, and Texas. Using these models, the number of total cases in the week after November 19th was predicted. The RMSE and R^2 were calculated and models were compared.

I think this project gave a good opportunity to explore real world data and practice skills of building models. It provided me with a good way to apply the knowledge I learned in class to the real world.

Yiyang Zhang:

What I did in the project included: participating to make decisions on the dataset to work on, coding for: random forest regressor and multi-layer perceptron and corresponding data visualizations, doing literature reviews, and the final report write-up.

I honestly think that this is a great opportunity to have hands-on experience on analyzing real world biomedical data using the skills we learned throughout the semester. I have learned a lot on how to deal with time series dataset and gained some precious experience of learning and coding to set up a multi-layer perceptron, a neural network that I have not learned before. To me, even though the final result might not be very optimal from a professional data scientist's perspective, it is still a great experience overall.

Jiajie Shen:
In the project, I have done the part of multi-layer perceptron regression model, it is a supervised learning algorithm. And it can learn a nonlinear function approximator for either regression or classification. It is a relatively accurate model for us to fit our COVID-19 dataset. I find that the regression plot showing the predicted cases and actual cases are fitted very well. And I have built up the interactive plot to show total cases in North Carolina. It is very helpful for us to evaluate the total COVID-19 cases prediction.

From my perspective, it is a wonderful experience for me to explore the real-world dataset to use the methods that we have learned in class. And I believe interactive plots are very helpful for us to see the results of the output. This project really helped me to understand how to use the python methods to solve real-world dataset problems.

6. Reference

https://machinelearningmastery.com/how-to-develop-multilayer-perceptron-models-for-time-series-forecasting/

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

github repository for this project:
https://github.com/sl679/823_final_project