Covid-19-Time-Series-Modeling

Public

main 1 Branch 0 Tags Go to file t Go to file Add file Code

Bella3s add files 928e427 · now 20 Commits

images	add files	now
pdfs	add files	now
.gitignore	Initial commit	3 weeks ago
README.md	README & small edits	6 hours ago
index.ipynb	small edits	1 minute ago
owid-covid-data...	add downloaded data	2 weeks ago
world_country_a...	add files	now

README

Abstract

This project creates a model that forecasts the number of deaths by Covid-19 given the historical data. The Covid-19 pandemic was detrimental in the number of people who perished from this disease. The idea behind this project is that if we can create a predictive model that will forecast the amount of harm, then public health officials can be better informed, advised, and empowered to mitigate future deaths. Specifically, this project focuses on a continent-wide evaluation, looking at Asia.

This project goes through a model iteration process, starting with a naive time series model, and iterating through ARIMA, multivariate ARIMA, Linear Regression, Facebook's Prophet, and multivariate prophet models. The project finalizes on the multivariate prophet model as the final model based on the evaluation metric, root mean squared error, and evaluates this model on the holdout test set. Lastly, the project offers a few recommendations to the proposed business and next steps for future projects.

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

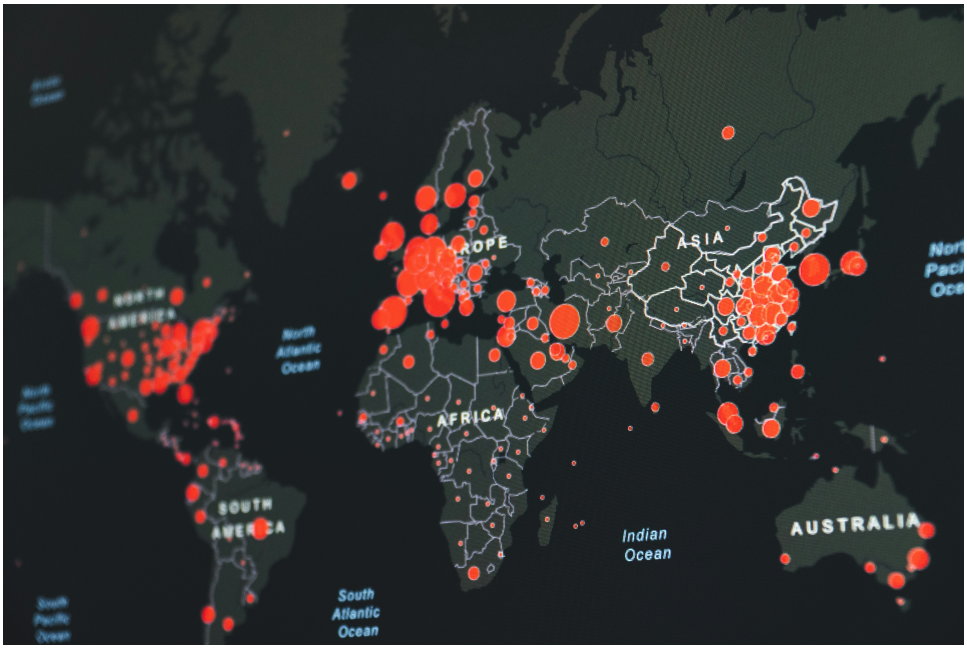
Languages

Jupyter Notebook 100.0%

Repository Structure

```
├── images ()
├── index.ipynb ()
├── owid-covid-data.csv ()
├── pdfs ()
├── README.md ()
├──
world_country_and_usa_states_latitude_and_longitude_values.csv
()
```

The Business + Business Problem



The Covid-19 pandemic needs little introduction as it left no corner of our world untouched. It devastated lives, and economies -- day-to-day life was dramatically altered for about 2-3 years. The specific business for this project is one akin to the World Health Organization (WHO) that 'treats data as a public good'(WHO Principles, but also wants to leverage this data to make recommendations to public health officials. The idea for this project is to create a model that will forecast how much worse the pandemic will be in order to advise public health officials and policy makers on how to best handle the situation such that damages are minimized.

The Data Source

The Data from this project is sourced directly from Kaggle, a dataset called Our World in Data - COVID-19. The company, Our World in Data combined their own data along with data from the John Hopkins University and the WHO. Further details on their data and sources can be found on the GitHub page here.

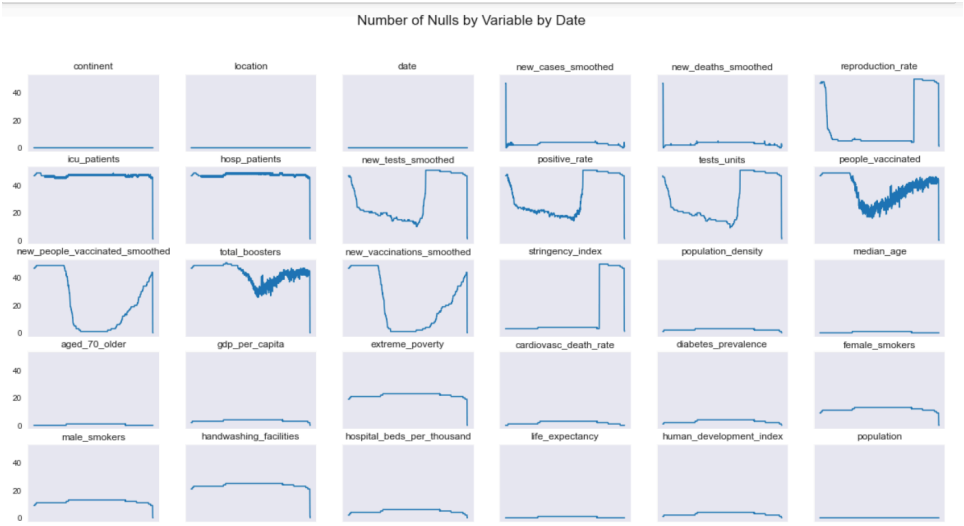
For visual purposes, the project also utilizes a dataset of latitude and longitude data of countries from Kaggle, originally sourced from google. See that data set here.

Data Exploration

In the initial data exploration, many of the variables are dropped as they are redundant for our purposes. The project looks at a visual of where the data is sourced from which leads to the decision to group the data by continent. This project focuses solely on the model created from the data in Asia.

Data Cleaning

There are numerous missing values in the complete dataset, as well as the Asia subset. The project looks at each of these and handles the missing values for each variable. This is done prior to the data being aggregated to one entry per day for the whole continent, but also while keeping mind to separate the train, validate, and test groups to avoid data leakage.

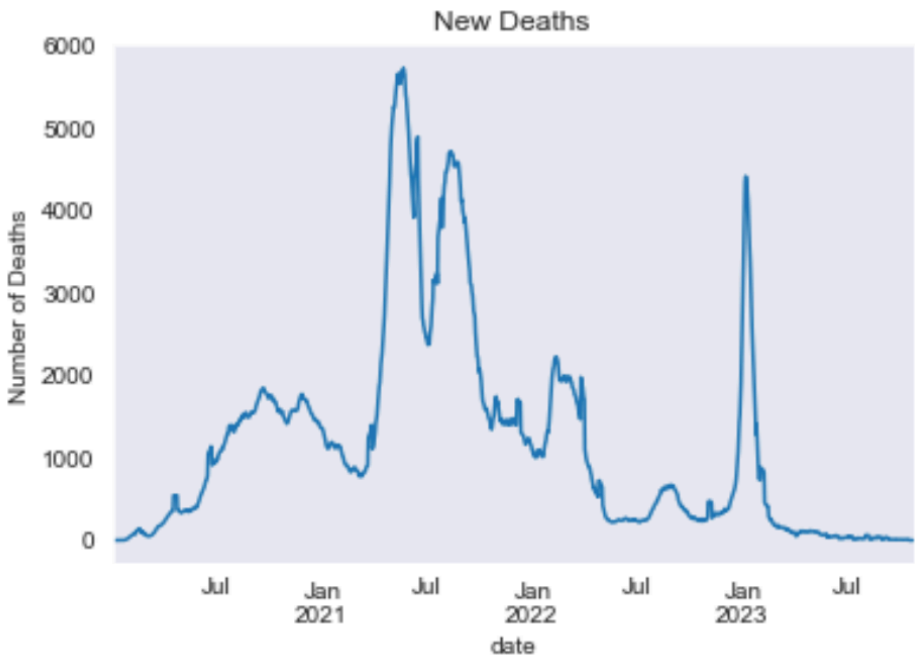


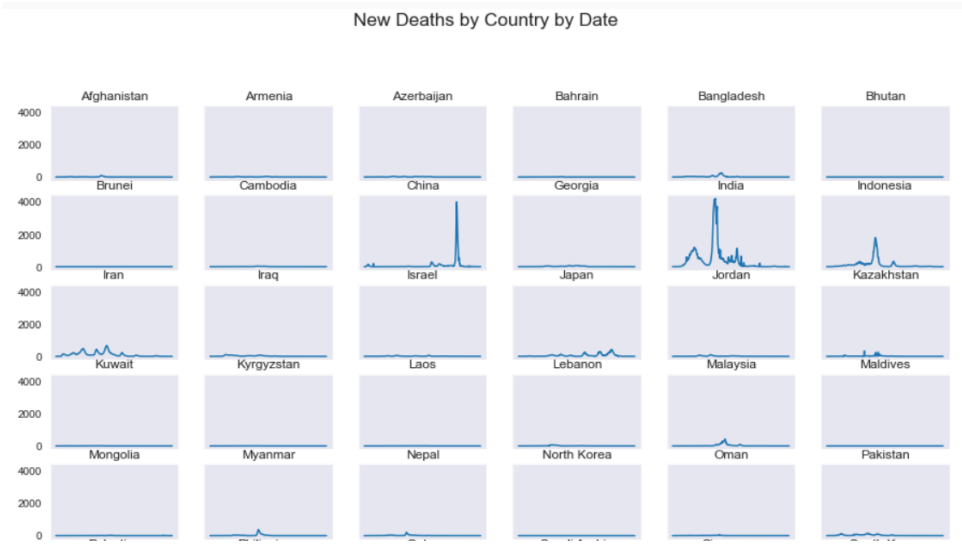
Aggregation + Final Exploration

After the data is aggregated (see table below), the project displays some visuals of the final dataset for a more thorough understanding of the data being used.

Variable	How to aggregate
`continent`	drop
`location`	drop
`date`	unique value per day
`new_cases_smoothed`	sum per day
`new_deaths_smoothed`	sum per day
`people_vaccinated`	sum per day (cumulative value)
`new_vaccinations_smoothed`	sum per day
`stringency_index`	average per day

Below is the target variable, the number of deaths per day, shown as a total aggregated value for the continent as well as by country.





Model Iteration

The models explored in this project include:

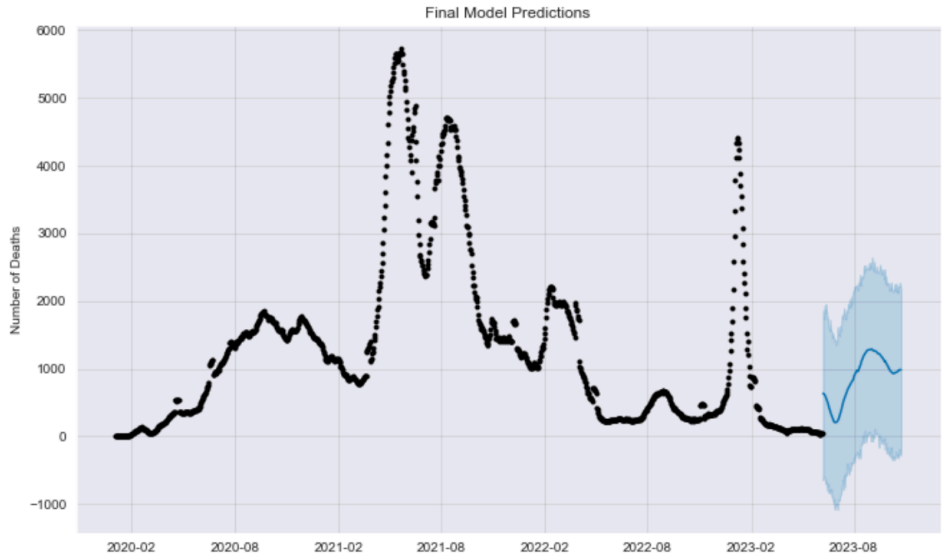
1. Naive Model
    - Our baseline model. The naive method uses 'today's' value as the prediction for 'tomorrow.'
2. ARIMA
    - Utilizing auto_arima to retrieve optimal p,d,q values
3. Multivariate ARIMA
    - Again utilizing auto_arima, but also including the other variables in our dataset, not just the target variable.
4. Linear Regression
5. Prophet
6. Multivariate Prophet

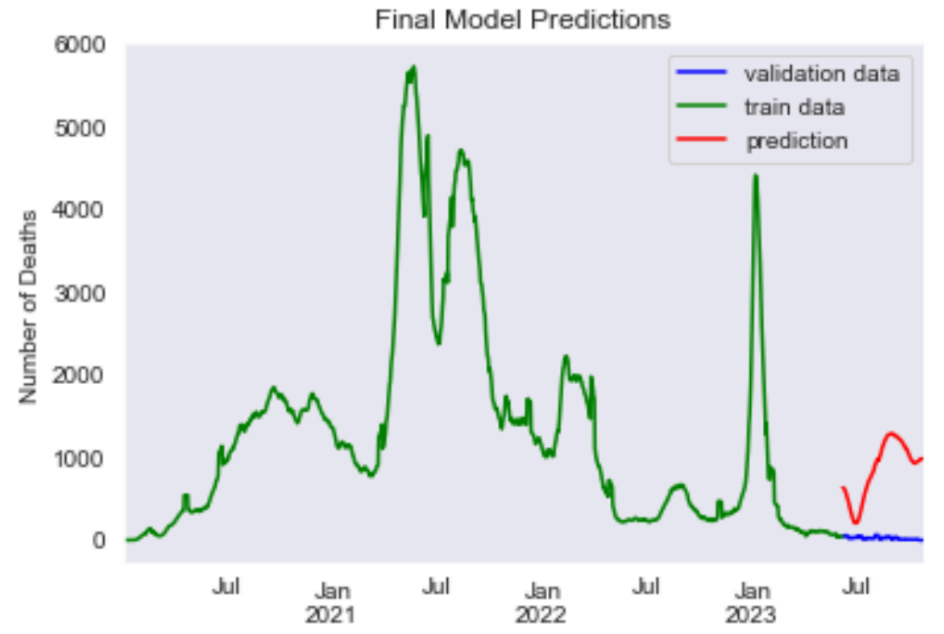
All Models were compared by their root mean squared error from the predictions on the validation set.

Model	RMSE
Naive	2,343
ARIMA	1,944
Multivariate ARIMA	968
Linear Regression	809
Prophet	2164
Multivariate Prophet	904

The Multivariate Prophet model was chosen as the final model even though the Linear Regression model had a lower RMSE because the Linear Regression model was predicting negative numbers for the number of deaths, the scores for these two models were quite close, and the prophet module is more suited for forecasting and analyzing trends in time series data.

Final Model Evaluation





Conclusion + Recommendations

Our model performed better on the validation data than it did on the test data. In part, this is because the pandemic had more or less ended by this time (mid 2023), which means the actual values for our data lie very close to zero, whereas our model is predicting the deaths to increase once again (similar to the actuals of July in 2020).

With an ending root mean squared error of 817,900 I would say this is a fairly good model, taking into account that this estimation is for the whole continent of Asia rather than one country alone.

1. **See how we can make our data more accurate.** There is always room for improvement in this area. The amount of missing data in the dataset that led to dropping so many variables is a prime example of how the data collection can be improved.

2. **Utilize models like these ones in future epidemics or pandemics.** Clearly, these predictive models can be quite useful. If we can forecast how detrimental an outbreak will be, then public health officials will be more equipped (and motivated) to handle the situation. They will be better prepared to implement policy changes to help curb the harmful affects of an outbreak, and ideally lessen the number of unnecessary deaths.

3. **Investigate how to further incorporate changing public health policies into time series data.** A big drawback of performing a time series model from this data was that interesting variables -- such as extreme poverty, diabetes prevalence, percent smokers, number of hospital beds per thousand people, population density, etc. -- were not viable to use in the time series prediction. The variables were constant for all dates which means they had no implication on the modeling process. However, these factors realistically play a role in the number of deaths caused by Covid-19. Furthermore, realistically, these numbers do change over time and thus could be used for the time series forecasting. If they can be more accurate and versatile per reporting period, then we can create a more proficient predictive model.

4. **Investigate why deaths tend to be reported higher on Fridays.** From both the multivariate prophet models (the final model and the model evaluated on the validation set), we can see a weekly trend where there is a spike in number of deaths on Fridays. It would be interesting to investigate why this is. Does it have to do with how the numbers are reported? We were using smoothed data so that seems less likely. Can we connect it to how and when people are contracting the disease? Can we connect it to situational occurrences in hospitals? Of course, to really dig into this phenomenon we would need to break the data up by country