Daily COVID Cases Prediction

Mingyi Gong, Danling Zhou, Yiqi Zhu
(Group: NO Work In Main Branch)

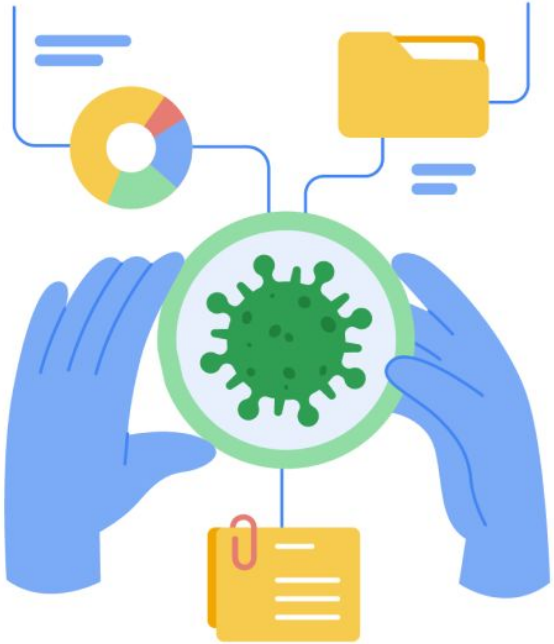
Project Overview

(i) Problem Description

- **Target Variable:** daily confirmed new cases

- **Independent Variables:** We will use variables including but not limited to demographics, economic info, policy, and vaccination data to predict our target variable.

- **Impact:** provide insight for future



Project Overview

(ii) Objectives

- **Prediction:** To make a machine learning model (or an ensemble model) to predict how the pandemic propagated.
- **Inference:** To interpret how economic conditions, government policies, and access to COVID vaccine impacted the spread of COVID.

(iii) Stakeholders

- Epidemiologists, public policy researchers, and the general public

Data Sources - Google Health COVID-19 Open Data Repository

(i) Overview

- It's one of the most comprehensive collections of up-to-date COVID-19-related information.
- Comprising data from more than 20,000 locations worldwide
- It contains a rich variety of data types to help public health professionals, researchers, policymakers and others in understanding and managing the virus.
- Last update date: September 15, 2022

https://health.google.com/covid-19/open-data/raw-data

Table	Indexed by[1]	Content	Source[2]	Download
Aggregated	`[key]` `[date]`	Flat, compressed table with records from (almost) all other tables joined by date and/or key; see below for more details	All tables below	.csv
Index	`[key]`	Various names and codes, useful for joining with other datasets	Wikidata, DataCommons, Eurostat	.csv .json
Demographics	`[key]`	Various (current[3]) population statistics	Wikidata, DataCommons, WorldBank, WorldPop, Eurostat	.csv .json
Economy	`[key]`	Various (current[3]) economic indicators	Wikidata, DataCommons, Eurostat	.csv .json
Epidemiology	`[key]` `[date]`	COVID-19 cases, deaths, recoveries and tests	Various[2]	.csv .json
Emergency declarations	`[key]` `[date]`	Government emergency declarations and mitigation policies	LawAtlas Project	.csv
Geography	`[key]`	Geographical information about the region	Wikidata	.csv .json
Health	`[key]`	Health indicators for the region	Wikidata, WorldBank, Eurostat	.csv .json
Hospitalizations	`[key]` `[date]`	Information related to patients of COVID-19 and hospitals	Various[2]	.csv .json
Mobility	`[key]` `[date]`	Various metrics related to the movement of people. *To download or use the data, you must agree to the Google Terms of Service*	Google	.csv .json

Data Sources - Google Health COVID-19 Open Data Repository

(ii) Dataset

- The original datasets contain a large number of missing values
- Manually aggregate datasets, joined by location and/or date
- Select columns with <20% missing values
- We expect that the final aggregated dataset will contain at least 100000 rows and 50 columns

```
epidemiology (12525825, 10)
demographics (21689, 19)
economy (404, 4)
emergency_declaration (8364, 104)
geography (22130, 8)
health (3504, 14)
hospitalization (1768485, 11)
mobility (6321226, 8)
search_trends (2713929, 424)
vaccination (2545118, 32)
government_response (303969, 22)
world_bank (215, 1405)
epi_by_age (3822577, 152)
epi_by_sex (3848340, 30)
key (22963, 15)
```

Methodology

(i) Data visualization

- Python data visualization libraries (matplotlib and seaborn)
- Focus on the time-course relationship

(ii) Missing data

- Select columns with <20% of missing values
- For the missing values, we will decide whether to use KNN to impute or drop the rows with too many missing values
- We recognize that imputing may lead to extra bias and dropping rows means letting go (possibly important) information
- Some boosting models can deal with missing data

Methodology

(iii) feature selection

- we expect to have a final table around 50 columns
- since we will be likely to use tree-based models heavily, feature selection is not a priority
- However, if needed, we will use stepwise selection methods and information gain

(iv) Modeling

- Linear regression
- Tree-based models (bagging and random forest)
- Boosting models (Gradient Boost, LGBM, XGBoost, CatBoost)
- Ensemble modeling
- Interested in attempting: neural networks

Conclusion

By analyzing the dataset derived from the Covid 19 Open Data, the project will provide us with a model (or an ensemble of models) that can predict the daily confirmed cases and insights on what impacted the transmission of COVID-19.

Any Questions?

Thanks for listening.

References

https://health.google.com/covid-19/open-data/raw-data