Insights on COVID-19 combined with Weather Data in The U.S

Akshara Reddy
Masters in Data Science
Illinois Institute of Technology
Chicago, United States
akudumula@hawk.iit.edu

Preethi Vempati
Masters in Artificial Intelligence
Illinois Institute of Technology
Chicago,United States
pvempati@hawk.iit.edu

Vishnubharathreddy Vankireddy
Masters in Data Science
Illinois Institute of Technology
Chicago,United States
vvankireddy@hawk.iit.edu

Harsha H. Chunduri
Masters in Data Science
Illinois Institute of Technology
Chicago, United States
hchunduri@hawk.iit.ed

***Abstract*— The COVID-19 pandemic had a significant effect on every human. The pandemic's results have had and continue to have a significant effect on our lives. We want to show various perspectives on how social health and weather influenced the movement of covid across different counties in the United States in the years 2020-2021 in this project. We wanted to see if factors such as temperature, social index, weekdays, and holidays, etc. played a role in the outbreak. To find these things, we have implemented a forecast model to predict the cases. The performance of the model was measured in terms of mean absolute error which is between bounds. This will help us in predicting the outcome of virus and help us to plan emergency health resources in a better way. The forecast model was trained on primary and contextual datasets to see if there is a change in the mean absolute error of the forecasted output.**

Keywords—COVID-19, Forecast, Mean Absolute Error, Social Vulnerability Index, Prophet, Performance measure, Exploratory Data Analysis

I. Introduction and Motivation

COVID-19, one of the deadliest virus outbreaks happened in Wuhan, China in the year 2019[1]. COVID-19 is an abbreviation for the corona virus disease which occurred in the year 2019 and hence the name COVID-19. This outbreak first evolved into an epidemic and soon into a pandemic affecting lives all over the world. We wanted to understand the effects of Weather, Socio-economic status of people, etc. on the spread of the virus. For this we have considered COVID dataset [2] from the New York Times' GitHub repository. The New York Times creates and uses data for reporting in their stories. Data is gathered from counties and municipal health departments in every state because local health agencies report data earlier than state authorities. This allows them to remain current and ahead of the competition. The data was made available to anyone who wanted to learn more about it. To understand the affects of contextual data, we have taken weather data from NCEI [3] and the social vulnerability dataset [4]. Weather variables for each state are often considered as an additional dataset to fully understand the impact of changing weather conditions. People in colder climates will spend a little less time outside. We are curious if COVID spread was reduced in these conditions relative to warmer climates. The National Centers for Environmental Information provides this dataset, which shows global surface temperatures for various nations (NCEI). There are no limits on how one wants to use this data for educational or research purposes. The dataset taken from the NCEI is checked thoroughly and random errors are removed from the NCEI end as the data is used by many scientists and researchers. We want to find out about people's socioeconomic status to figure out how the pandemic influenced various groups of people. We chose this because the upper middle class has greater access to food and can afford lockout healthcare services, while the working poor survive on a day-to-day basis and cannot afford to sit at home. This dataset is a 5-year estimate given by the American Community Survey (ACS) [5]. Since the COVID data [2] contains data in time series format, we experimented with the Facebook's forecasting model prophet [6-7].

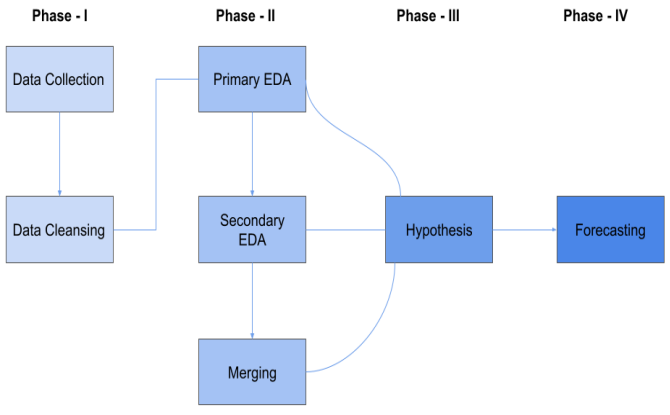


Fig1.Different Phases

The project was divided into four phases. In the primary phase, we collected the data and cleaned the data. In the second phase of the project, Primary dataset's EDA was performed i.e., EDA based on only the COVID dataset. Similarly, EDA was performed on the secondary dataset and then primary data was merged with the secondary data and EDA was performed on the merged dataset. In the first phase, different hypotheses tests were performed on different variables. In the final phase of the project, we forecasted the number of cases in the coming few months. The forecast was done individually on the covid dataset and then on the merged dataset.

II. Exploratory Data Analysis

A. *Primary Dataset Covid-19*

	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061.0	1	0.0
1	2020-01-22	Snohomish	Washington	53061.0	1	0.0
2	2020-01-23	Snohomish	Washington	53061.0	1	0.0
3	2020-01-24	Cook	Illinois	17031.0	1	0.0
4	2020-01-24	Snohomish	Washington	53061.0	1	0.0

Variables	Definition
date	YYYY-MM-DD
county	STRING
state	STRING
fips	UNIQUE NUMBER
cases	NUMBER
deaths	NUMBER

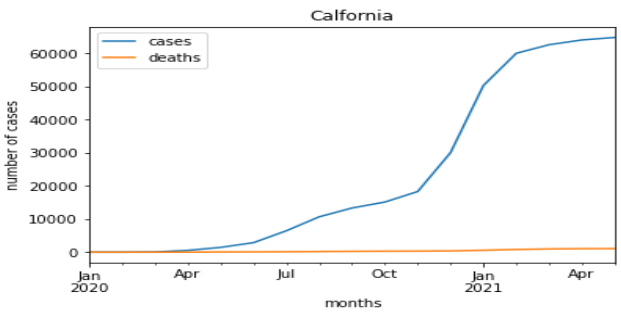


Fig2. Months vs number of cases in California

The above graph shows the graph with the number of cases on the y-axis and the corresponding months on the x-axis for the state of California. It can be seen that from November 2020, there is a steep increase in cases reported.

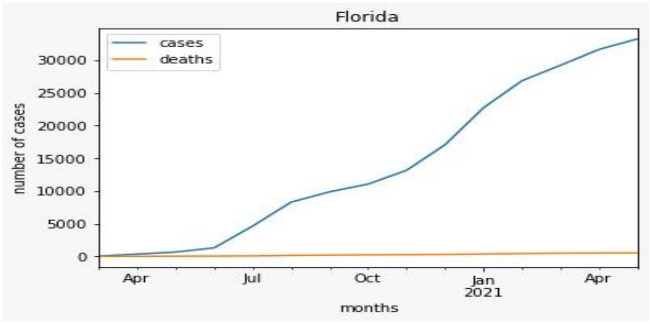


Fig3. Months vs number of cases in Florida

In Florida, there is a gradual increase in after November in the number of cases reported compared to the steep increase in California.

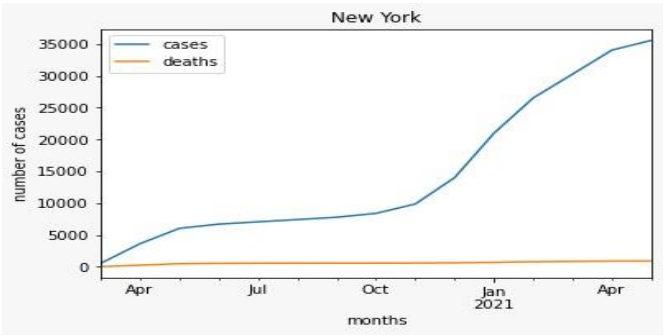


Fig4. Months vs number of cases in New York

For the state of New York, although there seems to be a steep increase in the number of cases reported, it is not as high as California.

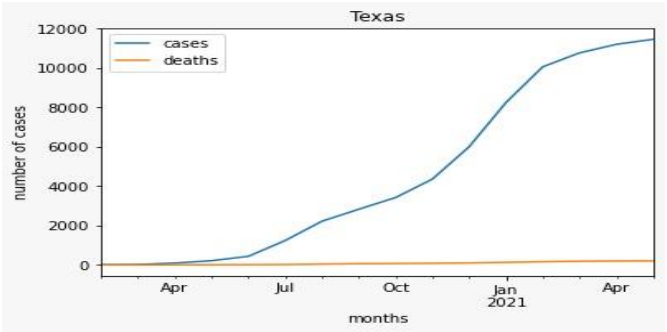


Fig5. Months vs number of cases in Texas

Above graph shows the number of cases and deaths reported for the state 'Texas'. The cases reported were highest during the months of October 2020 to Mid-February 2021.

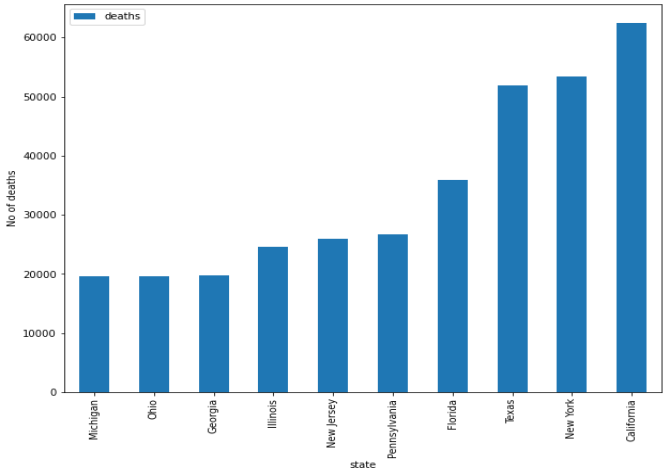


Fig6.states vs no of deaths

In the graph above, the top ten states which reported the highest number of deaths are shown. The states at the bottom, i.e., Michigan, Ohio and Georgia reported close to around 20000 deaths and are almost equal. Highest number of death reports were in California.

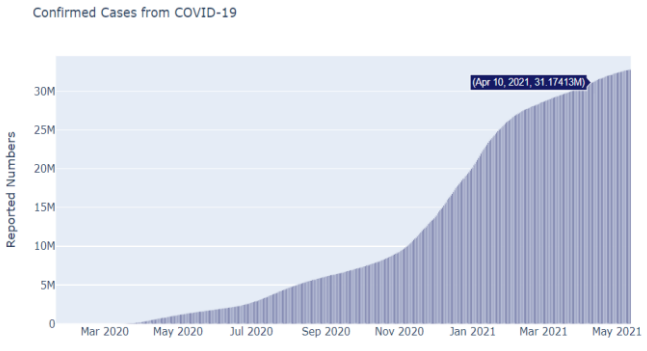


Fig7.confirmedcases in the USA

The above chart is an interactive chart which shows the reported number of positive COVID cases from the month of March 2020 to the month of May 2021.

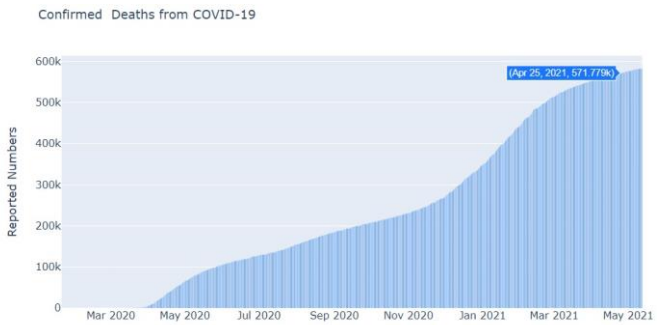


Fig8.Confirmeddeaths in the USA

This chart shows the number of deaths from the month of March 2020 till May 2021. The charts with confirmed cases and confirmed deaths look similar and it can be understood that as the number of cases rose, the number of deaths also rose.

Secondary Datasets:

```
svi_all_counties_cases_corr = svi_all_counties_cases[['confirmed_cases', 'EP_POV', 'EP_LIMENG', 'EP_AGE65', 'EP_MINRTY', 'EP
corr = svi_all_counties_cases_corr.corr()
table = corr.style.background_gradient(cmap='Blues').format('{0:,.3f}')
table
```

	confirmed_cases	EP_POV	EP_LIMENG	EP_AGE65	EP_MINRTY	EP_UNEMP	EP_NOHSDP	EP_AGE17	EP_DISABL	EP_SNGPNT	EP_MUNIT
confirmed_cases	1.000	-0.059	0.284	-0.199	0.219	0.019	-0.028	0.058	-0.208	0.084	0.426
EP_POV	-0.059	1.000	0.065	-0.105	0.448	0.648	0.626	0.076	0.482	0.517	-0.162
EP_LIMENG	0.284	0.065	1.000	-0.296	0.552	0.037	0.456	0.307	-0.223	0.193	0.222
EP_AGE65	-0.199	-0.105	-0.296	1.000	-0.396	-0.136	-0.133	-0.569	0.427	-0.462	-0.363
EP_MINRTY	0.219	0.448	0.552	-0.396	1.000	0.454	0.494	0.306	-0.079	0.564	0.184
EP_UNEMP	0.019	0.648	0.037	-0.136	0.454	1.000	0.445	0.066	0.403	0.469	-0.092
EP_NOHSDP	-0.028	0.626	0.456	-0.133	0.494	0.445	1.000	0.218	0.428	0.374	-0.298
EP_AGE17	0.058	0.076	0.307	-0.569	0.306	0.066	0.218	1.000	-0.279	0.421	-0.037
EP_DISABL	-0.208	0.482	-0.223	0.427	-0.079	0.403	0.428	-0.279	1.000	0.028	-0.450
EP_SNGPNT	0.084	0.517	0.193	-0.462	0.564	0.469	0.374	0.421	0.028	1.000	0.055
EP_MUNIT	0.426	-0.162	0.222	-0.363	0.184	-0.092	-0.298	-0.037	-0.450	0.055	1.000
EP_MOBILE	-0.198	0.513	-0.034	0.129	0.168	0.358	0.564	-0.057	0.541	0.094	-0.457
EP_CROWD	0.124	0.302	0.444	-0.327	0.519	0.339	0.357	0.434	-0.087	0.404	0.062
EP_NOVEH	0.078	0.463	0.028	-0.170	0.368	0.480	0.284	0.038	0.161	0.434	0.152
EP_GROUPQ	-0.060	0.173	0.019	-0.150	0.159	0.071	0.142	-0.330	0.053	0.055	0.004
RPL_THEMES	0.037	0.788	0.292	-0.205	0.619	0.650	0.736	0.188	0.475	0.612	-0.099

Fig9.Correlation for SVI data

The features such as Minority, Unemployment, total number of people in house, were highly correlated with the cases.

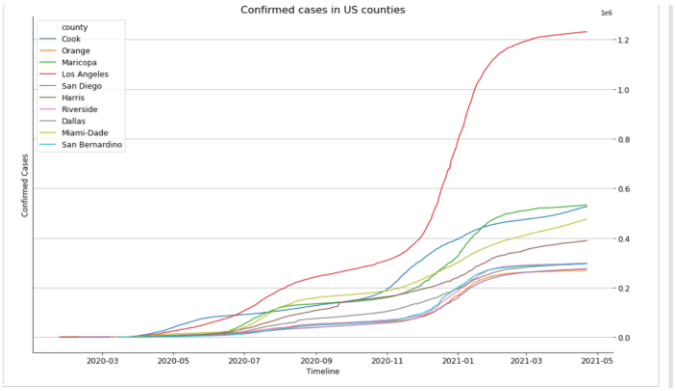


Fig10.Top 10 counties with most number of confirmed cases

Top 10 counties with the highest number of cases are displayed in the above chart. The highest cases were occurred in the Orange county in California. Cook county in the Illinois was the second highest reported county after California.

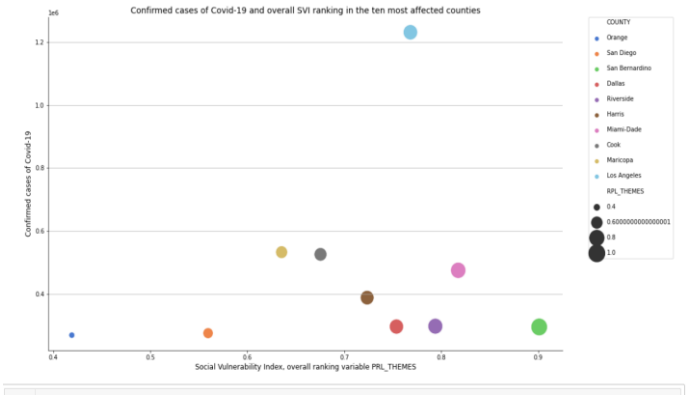


Fig11. Confirmed cases and overall SVI ranking in ten most affected counties.

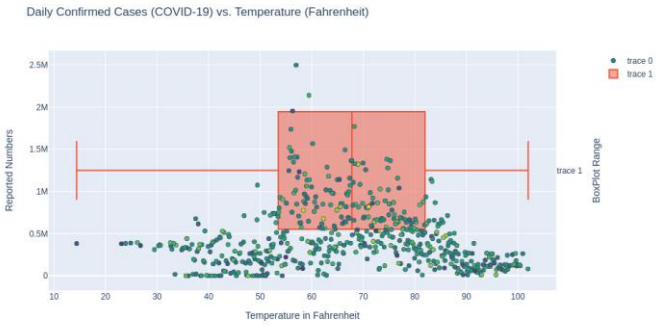


Fig12.Temperature vs Reported Cases.

The graph above shows the number of cases recorded with the temperature. Most cases fall in range between 54- and 85-degrees Fahrenheit. More numbers were seen to be reported. The reported numbers were in millions and the plot is showing confirmed cases against temperature.

III. MODELLING

The modelling is done as shown in the stages below.

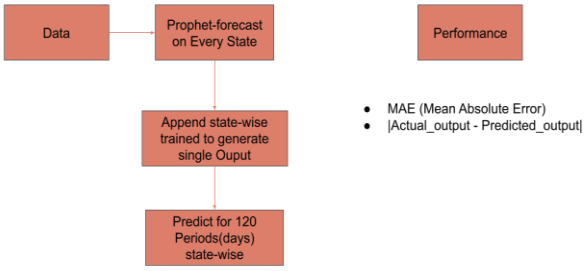


Fig13.Modelling Representation

First, data is split and passed on to the forecast model. To predict the number of cases likely to occur is always going to be the probable number of cases that might occur based on the events that have already occurred. For this particular use case, we decided to move forward with Facebook's "Prophet forecasting model"[1]. Prophet is developed by Facebook and is made open source to the public. This model is being used in the industry for various forecasting problems. This model is being used in various fields ranging from agriculture to banking, finance, and healthcare. However, it makes no sense to cross validate in the time series data as we cannot train future samples, that is, future dates with the present. So, to tackle this, validation time is taken for a certain period already recorded rather than cross validating. For the model, the date column is converted to datetime format. Then, we observed that every state had different numbers, some states high and some states low. So, the prophet time series was trained individually on every state and then appended everything back to generate an overall set but trained specific to the state.

Two cases were considered for the modelling part, one case for just covid dataset and the other one for combined or merged dataset. For the final modified dataset, for every state when merged the data is becoming too huge in more than 10's of GBs in size so for the final dataset, we didn't have much computational resource, so we took only the top five worst hit states that are "Illinois", "Arizona", "California", "Texas" and "Florida". However, for the initial dataset, all the states were considered.

Dates for the final predictions are a 5-month training, one month validation and 3–4-month prediction. The dates can be changed easily as we have put the dates in separate variables so as to ensure the code works smoothly when modified. Small note is that it should be in the formal YYYY-MM-DD.

Initially to measure the model, we decided to predict cases for the months of August and September 2020. It can be seen in the picture below.

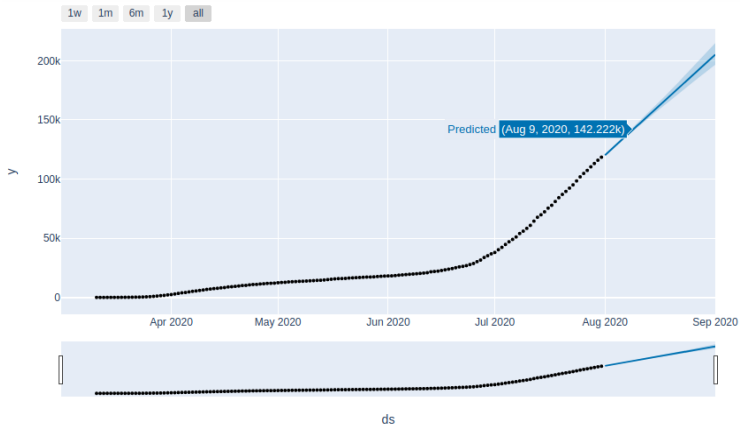


Fig14. Model forecast chart for the months of August and September for comparison purposes.

The forecast model was working without any issues and the predicted value was not too off which can be seen in the image below.

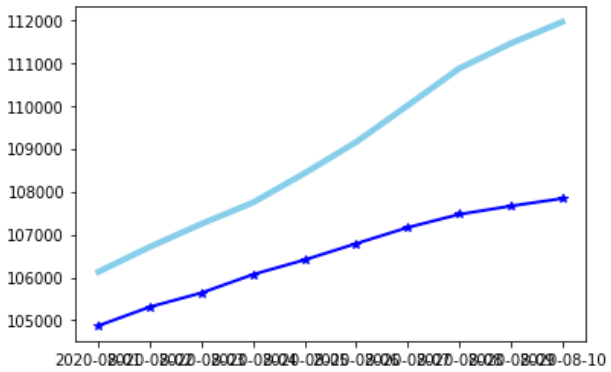


Fig15. Sky blue lines represents the predicted values whereas the dark blue lines represent the yhat_lower values.

There are four most important variables that are to be considered after the forecast model predictions are produced. They are as follows:

1. Ds (date)
2. Yhat (Predictor)
3. Yhat_upper (Upper boundary)
4. Yhat_lower (Lower boundary)

	ds	yhat	yhat_lower	yhat_upper	y	state
0	2020-08-01	105154.325642	104877.000360	105426.141479	106131	ILLINOIS
1	2020-08-02	105629.102417	105314.003330	105966.986727	106713	ILLINOIS
2	2020-08-03	105977.445746	105640.287192	106336.960101	107247	ILLINOIS
3	2020-08-04	106441.652063	106068.714643	106863.018308	107744	ILLINOIS
4	2020-08-05	106877.250411	106407.864493	107360.037804	108425	ILLINOIS
...	...	...	...	...	...	...

Above is a snippet of prediction set for all values in the dataset hence it is showing from August as we trained this on just the covid data to see the MAEs of prediction. yhat_upper and yhat_lower represent the interval in which the forecast lies, also it can be put as the "bounds of uncertainty" Uncertainty intervals are needed because generally there are three cases where uncertainty arises. Uncertainty in Trend, season, and noise [6-7]. We never know how trend changes, for example vaccine inoculations have started. The trend following this will change and a perfect prediction does not exist. So, these bounds are important to understand that the predictions could fall between either bound. The predicted cases in our case were higher than that of yhat_lower bound.

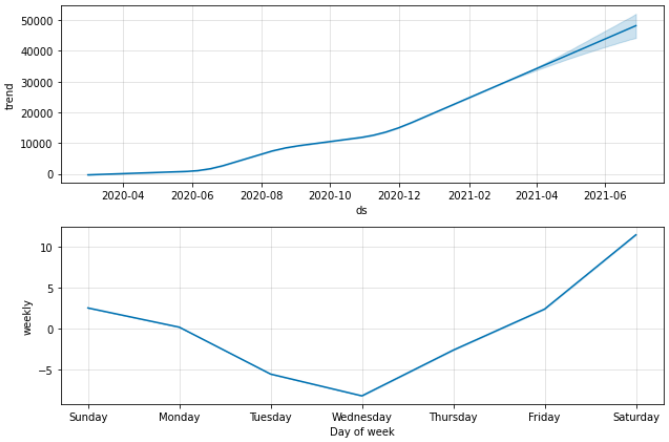


Fig16. Predictions without considering the holidays data.

One interesting thing about the prophet forecasting is that it has holiday-data functionality. We can see here that the spread is at the bottom in the middle of the week whereas it keeps rising till Saturday. But, when the holidays data in the U.S is considered, there is a decline. For different types of information, the forecast also changes.

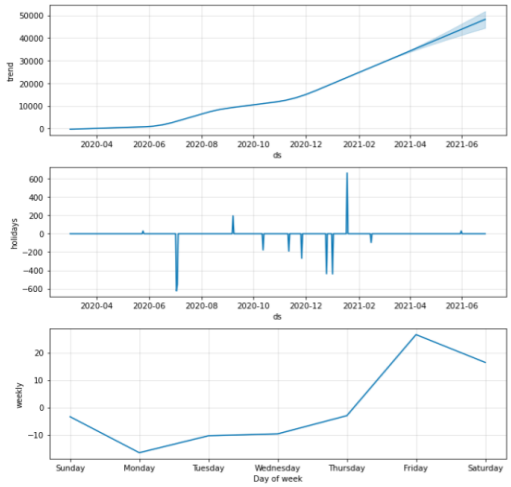


Fig17. Predictions with holidays data.

The performance measure for this forecasting is the mean absolute error which can be described as:

$$MAE = \sum_{i=1}^{n} |Y_i - X_i| \frac{1}{n}$$

Here,
Y_i = The predicted output
X_i = The actual output
n = Number of points in the data

	state	MAE
0	Washington	3532.853057
1	Illinois	5052.130077
2	California	23516.160976
3	Arizona	22110.389074
4	Massachusetts	15037.961903
5	Wisconsin	3619.521854
6	Texas	4443.012758
7	Nebraska	1094.590535
8	Utah	5082.533671
9	Oregon	1687.093401

The MAE of all the above states by taking only the COVID dataset is shown above. The MAE is always within the bounds of yhat_upper and yhat_lower.

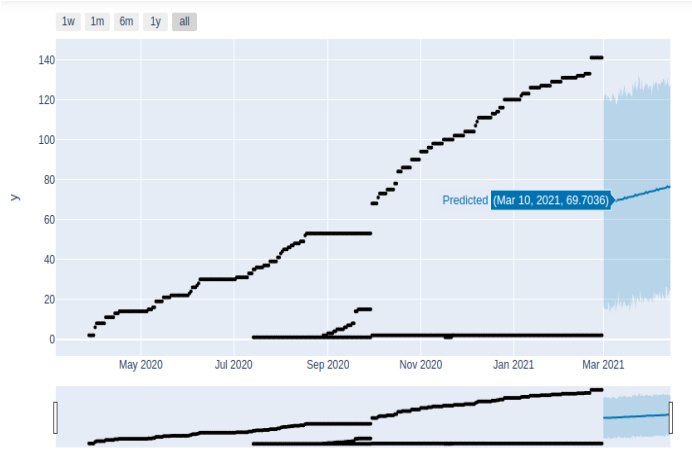


Fig18. Predicted values overall dataset for the 120 days.

The above chart shows the cases prediction based on overall dataset together with the contextual dataset. The graph overlapped because of a high training period. The model was trained from March 1, 2020 till December of 2021 with a validation set of 1 month and forecast for 120 days.
The performance measure of the mode can be seen below.

	state	MAE
0	Illinois	28415.970574
1	California	146758.412830
2	Florida	34120.257129
3	Texas	34334.241765
4	Arizona	56277.140195

The output was close to the lower boundary of the state.

A. *Abbreviations and Acronyms*

MAE – Mean Absolute Error

Yhat_upper – Target variable upper boundary

Yhat_lower – Target variable lower boundary

Y – Target Variable

ds – Date variable

B. *Equations*

$$\text{MAE} = \sum_{i=1}^{n} |Y_i - X_i| \frac{1}{n} \quad (1)$$

C. Conclusion and Future Scope

In this project, we have predicted the number of cases to be occurred in the next few days in the U.S with the help of covid, weather and social vulnerability datasets. The datasets can be found in appendix -A. Below are the limitations and future scope for the project.

- Patient demographics could not be taken which could have helped performance. Risk of running into heterogeneity as multiple datasets were combined.
- Model can be implemented easily to Counties, Countries, Cities effortlessly. Data of population inoculated with vaccine can be taken.
- Flu related information can be appropriately taken to differentiate covid with common flu.
- Sensory loss (Taste and Smell) data on a particular day with number of covid tests is one of the top predictors (doctor advice)
- Seasonal data with respect to every country can be added.

REFERENCES

[1] Zhu, Hengbo, Li Wei, and Ping Niu. "The novel coronavirus outbreak in Wuhan, China." Global health research and policy 5.1 (2020): 1-3.)

[2] https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

[3] https://www.ncei.noaa.gov/data/global-summary-of-the-day/archive/2020.tar.gz

[4] https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html

[5] https://www.census.gov/programs-surveys/acs/

[6] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1 (2018): 37-45.

[7] https://facebook.github.io/prophet/

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wort- man Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020

[9] https://facebook.github.io/prophet/

IV. APPENDIX

A. *Datasheets*

There are three different datasets that we have considered in this project. The datasheets for the same has been given below.

This document is taken from *Datasheets for Datasets* by Gebru *et al.* [8]. Hereafter, the covid dataset will be called "C", weather dataset will be called "W" and SVI dataset will

be called “S” and the datasheet answers for the same are written for there different datasets in the same section.

MOTIVATION

For what purpose was the dataset created?

This dataset was created by New York Times for reporting and tracking the covid-19 cases and deaths from the month of March 2020 till date. This dataset was used to report in the news.

New York Times

What support was needed to make this dataset?

Journalists, Healthcare workers, Medical department. State healthcare workers. Hundreds of journalists used to follow up regularly with healthcare workers on finding out the number of cases and deaths reported.

COMPOSITION

What do the instances that comprise the dataset represent?

This dataset represents county wise cases and deaths reported for every state under the United States in comma separated values. This document contains county information, fips code, cases reported in numbers. The entire dataset is text and number based.

How many instances are there in total (of each type, if appropriate)?

Dataset contains 1316501 instances and 6 variables.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The overall data from the GitHub repository will be updated every day for all the counties in every state within the United States.

What data does each instance consist of?

Data consists of date (YYYY-MM-DD), county (STRING), state(STRING), FIPS code(INTEGER UNIQUE), cases(NUMBER) and deaths(NUMBER).

Is there a label or target associated with each instance?

For every instance, we would like to predict the cases for upcoming months and this predictor will be the number of ‘cases’ which is our target variable.

Is any information missing from individual instances?

Initially, for the months of March 2020, not many cases were reported from outlying minor islands in the U.S and as a result not all counties were recorded in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)?

Yes, for predicting only on the COVID-19 dataset, splitting the data in the last month of prediction will be separated for validation and the rest of the dataset goes into model training. Validation is also in time series hence cross validation is not appropriate in our use case as we can’t mix future predictions with current reported cases randomly. Testing data is taken for 120 days.

Are there any errors, sources of noise, or redundancies in the dataset? The reported cases are aggregated every day, so the data looks ever increasing. The reported cases are always increasing. To find out present day cases, previous day cases must be subtracted.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

NO

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ non-public communications)?

Dataset is openly accessible by everyone who wishes to understand the data and use it for research purposes.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

NO. The dataset is not offensive or threatening in any way.

Does the dataset relate to people?

NO

COLLECTION

How was the data associated with each instance acquired?

Data is collected from counties, and regional health departments in every state as the data is reported earlier by the local health authorities compared to the state authorities. This helps them stay ahead and updated.

Over what timeframe was the data collected?

01-Jan-2020 till date i.e., 13-May-2021

What mechanisms or procedures were used to collect the data?

Calls and Records from journalists to local health authorities.

What was the resource cost of collecting the data?

Journalists were working day and night to find out the number of cases reported as some patients may go from state to state and in order to track, a lot of calls were to be done initially. Later, the health departments released bulletins and numbers, and this was taken by the journalists.

Who was involved in the data collection process?
Journalists and Healthcare Authorities

Is there a repository that links to any or all papers or systems that use the dataset?
https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

what tasks could the datasets be used for?
Predicting the next outbreak or next wave.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
The dataset collected is not in any way taking information such as gender, race, color, ethnicity, or religion of

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
New York Times Reporting Team

How can the owner/curator/manager of the dataset can be contacted?
covid-data@nytimes.com

Will older versions of the dataset continue to be supported/hosted/maintained?
YES. The dataset is time series aggregated.

PREPROCESSING/CLEANING

Was any preprocessing/cleaning/labeling of the data done?
The dataset was aggregated from previous cases, so the model was changed accordingly for this dataset.

Is the software used to preprocess/clean/label the instances available?
Python-3.8.5

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Reporting in News

How will the dataset be distributed?
GitHub: https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

MOTIVATION

For what purpose was the dataset created?
Every community must prepare for and respond to hazardous events, whether a natural disaster like a tornado or a disease outbreak, or an event such as a harmful chemical spill. So, the degree to which a community exhibits certain social conditions like high poverty, low percentage of vehicle access, or crowded households, may affect that community's ability to prevent human suffering and financial loss in the event of disaster.
Such factors describe a community's social vulnerability. The motivation is to investigate the correlation between Covid-19 cases and deaths in US counties and specific social vulnerability indicators. Based on features like unemployment, poverty, age, etc. US agency gives a social vulnerability index.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

ATSDR's Geospatial Research, Analysis & Services Program (GRASP) created Centers for Disease Control and Prevention Social Vulnerability Index.

What support was needed to make this dataset?
Census tracts, Journalists, Healthcare workers, Medical department, local planners.

COMPOSITION

What do the instances that comprise the dataset represent?
SVI provides specific socially and spatially relevant information to help public health officials and local planners. It has instances like Unemployment, age, minority status, poverty etc.

How many instances are there in total (of each type, if appropriate)?
Dataset contains 3142 instances and 123 variables.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The overall SVI data will be updated once a year after the census tract is done. This is not a sample of instances from a larger dataset.

What data does each instance consist of?
Data consists of instances like- Unemployment, Age, County, Minority status, Poverty, location, FIPS code (INTEGER UNIQUE), Income etc.

Is there a label or target associated with each instance?

For every instance, there is a predictor instance RPL_THEMES which is our target variable particularly for the SVI dataset. RPL_THEMES is an overall ranking for social indicators of each county.

Is any information missing from individual instances?

Initially, for the months of March 2020, not many cases were reported from outlying minor islands in the U.S and as a result not all counties were recorded in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)?

Yes, we can use this dataset for predicting the number of covid cases in a county based on social vulnerability indicators. By using confirmed cases as our target variable and social vulnerability indicators as x variables, we can split the data to train and test data using SVI dataset.

Are there any errors, sources of noise, or redundancies in the dataset?

Indicators such as Poverty, Unemployment looks as it is-highly correlated with confirmed cases but that is not always true.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

NO

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

Dataset is openly accessible by everyone who wishes to understand the data and use it for research purposes.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

NO. The dataset is not offensive or threatening in any way.

Does the dataset relate to people?

Yes, most of the variables in the dataset are related to a person's socio-economic status.

COLLECTION

How was the data associated with each instance acquired?

Data is collected from counties, local planners, census tracts and regional health departments in every state to help public health officials and emergency response planners identify and map the communities that will most likely need support before, during, and after a hazardous event. This helps them stay ahead and updated.

Over what timeframe was the data collected?

American Community Survey (ACS), 2014-2018 (5-year) data.

What mechanisms or procedures were used to collect the data?

For US-wide or multi-state mapping and analysis, we used the US database, in which all tracts are ranked against one another. For individual state mapping and analysis, use the state-specific database, in which tracts are ranked only against other tracts in the specified state.The vulnerability e index is created by counting the total number of flags in each census tract

Who was involved in the data collection process?

Us state or county authorities, as SVI data is mostly from US census tracts, US HHS and also ATSDR's Geospatial Research, Analysis & Services Program (GRASP),CDC.

What (other) tasks could the dataset be used for?

The dataset can also be used to understand the correlation between various social vulnerability indicators. For example, between poverty and unemployment or unemployment and minority status or age and unemployment.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

ATSDR's Geospatial Research, Analysis & Services Program (GRASP), CDC, US Government (census tracts).

How can the owner/curator/manager of the dataset be contacted?

We can contact the Centers for disease control and prevention of the US department of Health and human services**.**

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, past data can be used to understand how communities faced the calamities and the conditions they were in to overcome those hazardous events. Past data can also be used to predict what might happen in the future under certain conditions.

If others want to extend, augment, build on, contribute to the dataset, is there a mechanism for them to do so?

No, as the dataset is provided by the officials of US one cannot extend or augment the dataset without certain permissions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No, the dataset was not tampered as there were no null values or errors in the data. So, the analysis and preprocessing is done on the original dataset - SVI collected by the US government.

PREPROCESSING/CLEANING

Was any preprocessing/cleaning/labeling of the data done?
Preprocessing and cleaning of the data is done but there were no null values or errors in the dataset.

Is the software used to preprocess/clean/label the instances available?
Python-3.8.5

USES

Has the dataset been used for any tasks already?
Yes, the dataset was used to predict natural calamities, disease outbreak, chemical leak etc and help the community prepare how to face it even before the occurrence based on certain features.

How will the dataset be distributed?
https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

Datasheet for WEATHER dataset:

MOTIVATION

For what purpose was the dataset created?
This dataset was created by NCEI for recording the temperature, weather through sensors and is stored here. This dataset records nation wise, county wise and state wise overall temperature.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
National Centre for Environmental Information.

What support was needed to make this dataset?
Weather related temperature sensors located at land-based stations and satellite information.

COMPOSITION

What do the instances that comprise the dataset represent?
The instances in the data have information related to station number, date recorded, latitude and longitude, elevation, name of airport, temperature, dewpoint, sea level pressure, visibility, precipitation, gust and storm activities.

How many instances are there in total (of each type, if appropriate)?
Over 10 million instances are present, however we skimmed it down to the states of “California”, “Illinois”, “New York”, “Texas” and “Arizona”.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
We took instances from an overall dataset.

What data does each instance consist of?

Station	Date	Latitude	Longitude	Name
Number	MM/DD /YYYY	FLOAT	FLOAT	STRIN G

Rest of the attributes are all Float values.

Is there a label or target associated with each instance?

The target variable is taken from the COVID dataset mentioned in Appendix-A.

Is any information missing from individual instances?
Missing information is present, and it is represented by the “999.9” value.

Are there recommended data splits (e.g., training, development/validation, testing)?
Yes, the data is split with respect to the COVID dataset dates.

Are there any errors, sources of noise, or redundancies in the dataset?
Redundancy in data is given when there is a join happening between the covid and weather dataset as for every date, all the attributes get copied.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
NO

Does the dataset contain data that might be considered confidential?
Dataset is open to be accessible by everyone who wishes to use it for research purposes.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
NO. The dataset is not offensive or threatening in any way.

Does the dataset relate to people?
NO

COLLECTION

How was the data associated with each instance acquired?

Data is collected from satellites and weather-related sensors typically located in airports by the NECI.

Over what timeframe was the data collected?
01-Jan-2020 till date i.e., 13-May-2021.

What mechanisms or procedures were used to collect the data?
Data is directly downloaded from the tar files. An API functionality is also available.

Who was involved in the data collection process?
NCEI

Is there a repository that links to any or all papers or systems that use the dataset?
https://www.ncei.noaa.gov/data/global-summary-of-the-day/archive/2020.tar.gz

What (other) tasks could the dataset be used for?
Predicting the weather for various counties throughout the year. Farmers can use this to plan crops.

PREPROCESSING/CLEANING

Was any preprocessing/cleaning/labeling of the data done?
The dataset was joined based on station numbers by manually looking for the nearest airport to the county variable present in the COVID dataset.

Is the software used to preprocess/clean/label the instances available?
Python-3.8.5

USES

Has the dataset been used for any tasks already?
Yes, for various purposes.

How will the dataset be distributed?
Can be taken from https://www.ncei.noaa.gov/archive

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
NCEI

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
ncei.info@noaa.gov

B. Model Card

Model Card - Covid Cases Forecasting

Model Details:
- Developed by fakebook's data science research team.
- Published in the paper "Forecasting at Scale" in the year 2017.
- Time series forecasting

Intended Use:
- Fast predictions for time series data
- Agriculture, weather forecasting, market analysis, market forecasting,
- Open sourced by Facebook

Factors:
- Evaluation of the model can be different for different implementations.
- Depends on date or time series information.
- The formats of date and time should be converted to the format accepted by the model which is 'YYYY-MM-DD'

Metrics:
- Mean absolute error (Taken)
- Mean percentage error.

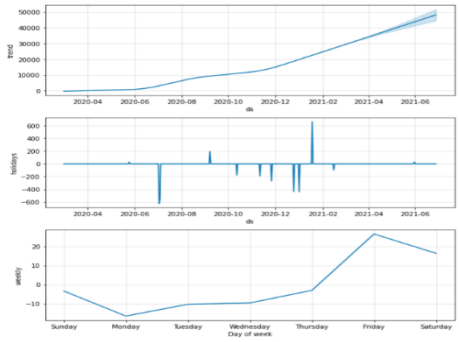
Training Data:
- For only the covid mode, we have done twice, one for time period Jan - July 2020 and predicted for August and September 2020.
- For the overall dataset, we have predicted for the months of November and December as computing resource was not handy
- Although we did use validation sets for training and predicted for 120 days starting from March 2021. Training data - March to December 2020, Validation - January and February and prediction for March, April, May and June
- No cross validation as future cases can't be mixed in this particular forecast

Evaluating data:
- Time series data with cases as predictor/target variable.
- Four months data in two different sets.

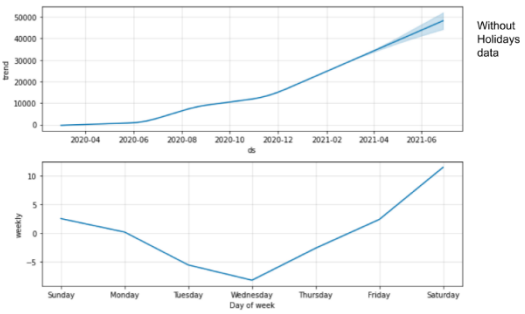
Ethical Considerations:
- No racial data is taken in this dataset.

Recommendations:
- Using holiday data will help predict cases better in a week
- Vaccination information can be taken.

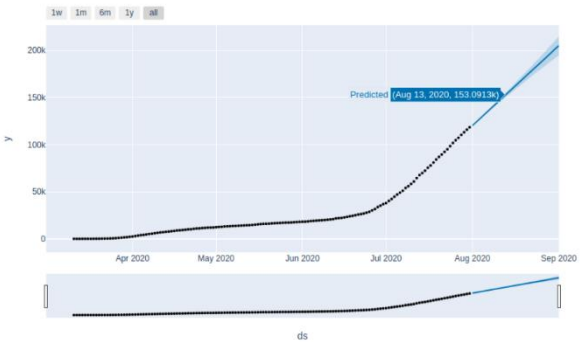
Model Predictions:
- With holiday data

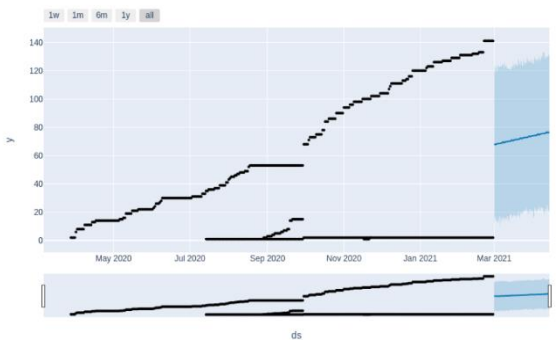


- Without holiday data



- Predictions





C. **Feedback Response**:

1. What technique(s) did they employ to code the more complex categorical variables like counties?

A. We did use FIPS code for combining one categorical variable but for the weather dataset we have used station number. Coding wise, we have merged the datasets based on FIPS and STATION numbers.

2. For the SVI dataset, which variables (or themes) had stronger positive correlations with Covid cases?

A. The temperature variable had a positive correlation with the increase in cases. The relation can be seen in Fig.12

3. What is the benefit of a country-wide model when it was previously determined that the incidence of Covid-19 cases positively correlates with population density? How applicable is the resulting model at the city/county level?

A. We have not used a country wide model as the population varies state by state. Instead, what we did was we have created a model for every state and then aggregated the output into one single data frame. The county level information was more crucial instead of a city wise model because a country wise model can help consider the areas surrounding the city as it can be taken in account as a precaution or for emergency measures.

4. How to set the user-defined holiday parameters in prophet?

A. The holiday parameters if needed in the U.S can be taken directly from the model fit function. A user defined parameter can be set as follows:

```
m = Prophet(holidays=holidays_dataframe)
forecast = m.fit(df). predict(future)
```

5. How did they choose which features are appropriate for the combined dataset?

A. We have checked the correlation between variables but decided that our use case related features be drawn. We have manually checked more than 100 variables, skimmed through the definitions and chose what was appropriate for our data. The information can be seen in the datasheets of weather and SVI.

6. How did you validate your predictions?

A. We have taken a one-month covid data for validation. Since this is a time series data, the validation is cross-validation but rather sequential. For ex: Mar-Oct(training), Nov (Validation set) and Dec (Prediction)

7. How did you validate your predictions?

A. It was hard to consider accuracy as the predictions might always tip over. For this reason, we have considered the mean absolute error with upper and lower bounds. The bounds are there to maximize our prediction confidence. The predictions were lower than the upper and higher than the lower bounds.

8. Does the model performance vary after the addition of weather and SVI data? How?

A. The performance of the model did vary slightly with some states giving better predictions with less mean absolute error while the states with less data had higher margins of error. Overall, the performance of the model slightly improved.

9. How well does the model perform on predictions on the COVID case count today?

A. As of 07-May, the predictions for California found to be 2.5K off the mark. The reason could be because of the vaccinations. We have not considered the vaccination data and we are sure that this produces a better result.

D. INDIVIDUAL REPORT

Contribution:

- Collecting the datasets for secondary data was quite an effort. I had to understand how to fit the secondary data so that I can make something useful out of the dataset.
- Particularly the weather dataset as the information available was not really in a good structure. Had to browse many sites and go through many datasets and check which data made more sense.
- I had taken the job of modelling for the cases forecasting. I used Facebook's prophet forecasting model to predict the number of cases to be reported in the upcoming months both for the primary and final (merged) datasets.
- Helping my fellow members with blockers.

Challenges:

- Learning about forecasting was quite a challenge as I never worked on time series data before.
- Understanding the datasets after merging took a toll on time to see if the time series models performed well.
- Performance metric is not straight forward and had to calculate the mean absolute error.
- Model issues, initially when the pandemic began, there weren't many cases recorded in every state and the model was not accepting the same data for every state as some states have less than 2 records. I have then changed the model to be trained from march.
- Preparing model for every state since countrywide model is not applicable. A loop was given to the model and the aggregations for every state model was appended and then created a single data frame for all the outputs.
- Holiday functionality for the model. Taking holidays as an input to the model and re iterates the performance of the model.
- Dividing the work was also a challenge as the project is a sequential one but I have divided it into phases and according to the skillset of the group.

10. Did you compare the trend predicted with the actual number of cases of covid? how did it perform?

A. As of 07-May, the predictions for California found to be 2.5K off the mark. The reason could be because of the vaccinations. We have not considered the vaccination data and we are sure that this produces a better result.

- Collaborating virtually was a lot tougher as everyone has different environments (python) to work, and version issues were persistent initially. Planning the virtual environment for python was key.

Learnings:

- I have always wanted to learn about the time series forecasting and with this project I have successfully learnt the majority of the time series forecasting in terms of modeling.
- I have understood the importance of contextual data for inclusion with datasets. Working on multiple datasets made me think in a holistic way.
- I have learnt about the boundaries of forecasting models which are really important to know as values can never be certain and a margin of error always occurs due to unforeseen circumstances.
- "Performance is not always about accuracy", this is one of the most important things I have learnt.
- Missing information was degrading the performance of the model and feature importance was key. Understanding the features by a correlation matrix was okay but not good enough so I had to manually read the definitions of the variables and include. This was really important in terms of learning.
- Learning about the missing values was quite hard and rewarding. Most of the time when working with curated datasets, there is no issue of missing values but working on raw data and preprocessing the data was good.

Group Performance:

- Merging the datasets was one of the key challenges.
- Virtual python environment and packages versioning as everyone was working on different versions of the software, integrating it into a one whole package was tough.
- Multiple datasets addition and manual evaluation of variables was challenging as the definitions of the variables were to be understood before integrating.

Grades:

- Akshara - A
    - Understanding the SVI dataset and doing the EDA was really helpful to get an understanding of the cumulative dataset forehand even before the model was built.
- Preethi - A
    - Was good with the primary EDA i.e., covid dataset and since the data is aggregated output of the number of cases functionality and EDA had to be very clever in terms of evaluation.
- Vishnu - A
    - Merging the datasets was quite a challenging task and vishnu did most of the merging part. Having to browse and evaluate joining variables was complex with limited resources. He stood up to the challenge and had to manually look for station numbers of nearest airports for merging.