Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
CHAPTER 5

Using interpretable machine learning identify factors contributing to COVID-19 cases in the United States

Ashish Garg
Walmart Global Tech, Bentonville, AR, United States

1 Introduction

1.1 Background of the study

In December 2019, in Wuhan, Hubei, China, several people who visited the wet seafood market were found infected with an unknown virus and the pathogen was similar to coronavirus (CoV). This unknown virus was named 2019 novel coronavirus or 2019-nCoV, which was later renamed by World Health Organization (WHO) as COVID-19 on 11th February 2020 (Coronavirus (COVID-19) events as they happen, n.d.; Sun et al., 2020). Fig. 1—COVID-19 Expansion shows the elaboration of COVID-19, which states what it stands for.

International Committee on Taxonomy of Viruses—Coronaviridae Study Group (ICTV-CSG) of Viruses named COVID-19 as SARS-CoV-2 based on naming nomenclature (Gorbalenya et al., 2020). Fig. 2—Coronavirus naming during the three major outbreaks shows the evolution of the names of different respiratory pandemics.

With the rate at which this virus was spreading, on 30th January 2020, WHO declared this as “Public Health Emergency of International Concern” and later on 11th March 2020 declared this as “pandemic” (Coronavirus (COVID-19) events as they happen, n.d.).

Coronaviruses are genetically diverse viruses found in animals and humans, which cause intestinal and respiratory infections. These were first identified in, 2002–03 in Guangdong Province, which was named as Severe Acute Respiratory Syndrome (SARS). This virus spread over 26 countries, infecting 8096 people with a fatality rate of 10% (WHO | Summary of probable SARS cases with onset of illness from 1st November 2002 to 31st July
It was later found that SARS originated from bats and was transmitted to humans through Himalayan palm civets or raccoon dogs. Another type of virus that originated from an animal is Middle East Respiratory Syndrome (MERS). This virus infected 2494 people with a fatality rate of 35% (WHO | Middle East respiratory syndrome coronavirus (MERS-CoV), 2020).

As of 24:00 PST on 29th August 2020, we have a total of 24,587,513 cases worldwide with 833,556 deaths with the maximum number of confirmed cases in the United States. Fig. 3—World dashboard for COVID-19 shows the spread of cases throughout the world as of 29th August 2020.
Right now we are months away from vaccines and need to use Non-pharmaceutical Interventions (NPIs) to curb the spread of the virus till we have vaccines available (Coronavirus Vaccine Tracker—The New York Times, n.d.; Nonpharmaceutical Interventions (NPIs) | CDC, n.d.). With the option to impose multiple NPIs it becomes difficult to choose which NPI/s will suppress new confirmed cases and sustain the same. Imposing a wrong NPI/s won’t impact new confirmed cases but moreover will adversely impact the economy and further create panic in people. To solve this problem, we need to understand mobility patterns of people against new confirmed cases, and this will help in drawing the inference of which mobility pattern spikes up the number of cases based on which an educated decision of imposing NPIs can be considered. A well decided imposed NPI/s will suppress current confirmed cases and sustain the same until it is in effect. This will further help in controlling peak healthcare demand and managing resources effectively.

1.2 Problem statement
COVID-19 is the worst pandemic that the world has experienced in the recent decade. It has claimed more than 800 thousand lives worldwide, with a close 25 million confirmed cases as of 29th August 2020 (WHO coronavirus disease (COVID-19) dashboard | WHO Coronavirus Disease (COVID-19) Dashboard, n.d.). It is on a sharp rise and is becoming more severe with every passing day. The United States is the worst hit by this pandemic, having more than half of total confirmed cases worldwide. 

As of 29th August 2020, we stand nearly the past 5 months since COVID-19 was termed as a pandemic by the World Health Organization (WHO) with no identified treatment or vaccine available for the same (Coronavirus (COVID-19) events as they happen, n.d.). Several health organizations are rushing towards the development of vaccines and are trying to expedite the development process, but still, we stand months away from the same (Coronavirus Vaccine Tracker—The New York Times, n.d.). The reason being several complications, regulations, and testing around the vaccine. A normal vaccine typically takes 18 months from inception to first delivery, even though with the expedition of process, time can be slightly reduced but not drastically. Once the vaccine is identified, mass production and availability for the public will again have an added timeframe. COVID-19 spreads from person to person through aerosols or droplets, which can be transmitted through contact routes or respiratory droplets.
Hence, until the vaccine gets available for the public, this pandemic needs to be controlled using other ways known as Non-pharmaceutical Interventions (NPIs).

Non-pharmaceutical Interventions (NPIs) are behavioral restrictions that are adhered to by an individual and as a society to curb the spread of the virus \((\text{Nonpharmaceutical Interventions (NPIs)} \mid \text{CDC, n.d.})\). Several countries have imposed such NPIs in the form of Social Distancing, closing down public places, schools, colleges, workplaces, etc. Imposing NPIs shows a positive effect on curbing the spread of the virus but harms the economy drastically. Hence, NPIs should be implemented to make maximum impact in curbing virus and make the least impact on the economy. Understanding of public mobility patterns with population density and confirmed cases will help in the understanding of the contribution of a place in the spread of COVID-19. This will further help in deciding which particular NPIs should be imposed, which will reduce interaction in that place leading to the curb of the spread of the virus.

Based on this thought, this study focuses on finding NPIs that should be imposed in the United States to curb the spread of COVID-19 in the absence of vaccines based on mobility patterns of public, population data, confirmed cases, and intensity of spread of COVID-19 in a particular region.

1.3 Aim and objectives

The primary aim of this study is to help make the data-backed decision on which NPIs should be imposed to curb the spread of COVID-19, which can be achieved by understanding the impact of the global mobility pattern of people along with population and population density in a region against new confirmed COVID-19 cases observed.

Based on this, the objectives can be

- To evaluate pattern and relationship between mobility factors, population, population density, and new active cases.
- To evaluate the performance of different machine learning models in predicting the number of active cases on a particular day.
- To deduce the interpretability of the model in predicting new active cases
- To identify features that have a high impact on new active cases.
- To identify the statistical model which is efficient in predicting new confirmed cases.
1.4 Research questions

Following research questions have been identified for this study.

- How does the different activity of people in places such as retail stores, parks, shopping center, residential area, etc. impact new confirmed COVID-19 cases.
- Is there any correlation of new confirmed COVID-19 cases with population and population density? COVID-19 has a high R₀, which means a high reproduction rate, so does places with a high population or population density are more prone to have a high number of COVID-19 cases.
- Do existing confirmed cases affect new confirmed COVID-19 cases. It has been found COVID-19 is spread from humans to humans, but we want to find out statistically there is any relationship between new confirmed cases with existing confirmed cases.
- Which features impact most new confirmed COVID-19 cases? This will help in determining which NPIs should be imposed to curb the spread of COVID-19.
- Which statistical model best predicts new confirmed COVID-19 cases using mobility features, population, and population density.

1.5 Significance of the study

- The research will explore how population intensity in different public places impacts new confirmed COVID-19 cases. These findings will help in deciding on which Nonpharmaceutical Interventions (NPIs) should be imposed to curb the spread of COVID-19.
- This research will also focus on understanding the impact of population and population density on new confirmed COVID-19 cases. This will help in understanding how prone a state is to have a severe COVID-19 outbreak.
- Another part of the study is to predict the number of new cases on a given day, and this will help in better planning and management of health services. We will also compare different models and their prediction efficiency to bring a contrast between models and identify the best model for this research.

1.6 Structure of the study

In this research, we follow a sequential approach. We start with “WHY” do we need to perform this research aiming at finding factors contributing to
the spread of COVID-19. It starts with the background of the study and dives into the problem statement, followed by the aim, objectives, and research questions.

Next in Section 2, we dive into reviewing existing literature that considers similar interests with COVID-19 and understands the point of view of other researchers. In Sections 2.1 and 2.2, we discuss different attributes of COVID-19 that help in understanding it better, such as “Basic Reproduction Number,” which explains how the rate at which infection is spreading, and the “Incubation Period” when the person infected starts showing symptoms. Post that in Section 2.3, we dive into where we stand with treatments and vaccines for the same, also understand the vaccine lifecycle. With this in Section 2.4, we also dive into other measures that can be leveraged in the absence of a vaccine to curb the spread, known as Non-pharmaceutical Interventions (NPIs). Further in Section 2.5, we discuss the strategy for implementing NPIs and Contact Tracing by different countries and evaluate their effect on COVID-19 as well as other socio-economic aspects.

Section 3 is about “HOW” we are going to conduct our research. We discuss different statistical models for this research along with evaluation metrics and steps to obtain model interpretability to understand the magnitude of the impact of features on new confirmed COVID-19 cases.

In Section 4 we evaluate experiments and results, defining the best model and further understanding from model interpretability. Finally, in Section 5 conclusion and future work are explained.

2 Related work

COVID-19, which has been declared a pandemic, is still on a streak of spread, with active cases multiplying every passing day. In this section, we will focus on elaborating on the dynamics of COVID-19, the current state of vaccine development, measures that can help curb spread in the absence of a vaccine, and the type of data that we are considering performing this research.

Fig. 4—Daily new confirmed COVID-19 cases above show the trend of new confirmed cases in different countries; currently, testing is implemented in a full-blown method, but still, it is limited. The number of confirmed cases is directly proportional to testing and actual cases may be more than confirmed cases (Moustakas, 2020). We can see in the given comparison the United States tops the chart with approximately the same high number
of confirmed cases from the past 60 days. Community spread has been confirmed in the United States, with New York having the worst outbreak.

COVID-19 virus can spread through respiratory droplets, which can pass from one person to another through verbal interaction, cough, or sneeze when the distance between the 2 is less than 1 m. Another way of transmission is contaminated surface to hands and then to nose, mouth, or eyes passing virus (Ong et al., 2020). Most of the countries have implemented various measures to curb the spread of the virus in the absence of vaccines which all aim towards reducing contact between people resulting in reducing the spread of the virus. These measures have been implemented in the form of Social Distancing, Lockdowns, Ban on public/social gatherings, workplace, school, college closure. These measures have positively impacted reducing the new number of confirmed cases, but a spike is always observed whenever these measures are relaxed.

2.1 Basic reproduction number (R₀)

COVID-19 confirmed cases are increasing exponentially throughout the world, which is possible due to the high Basic Reproduction Rate. Basic Reproduction Number (R₀ pronounced as R naught) is used to measure
the transmission dynamics of infections, viruses, and parasites. It is an epidemiologic metric with a numerical value, the higher the value more the virus has an affinity in spreading (Delamater, Street, Leslie, Yang, & Jacobsen, 2019; Dietz, 1993; van den Driessche & Watmough, 2008). Researches following Table 1 R0 estimates from different published papers demonstrate R0 calculated by different researchers using different statistical methods. Interestingly researchers have defined the range of R0 to be in 1.5–6.49 with a mean at 3.28, which is way higher than the R0 range provided by WHO of 1.4–2.5 with a mean at 1.95.

Table 1 R0 estimates from different published papers (Liu, Gayle, Wilder-Smith, & Rockløv, 2020).

| Study | Location | Study date | Methods | R0 estimates |
|-------|----------|------------|---------|--------------|
| Wu, Leung, and Leung (2020) | Wuhan | 31st December 2019–28th January 2020 | Stochastic Markov Chain Monte Carlo methods (MCMC) | 2.68 |
| Shen, Peng, Xiao, and Zhang (2020) | Hubei province | 12th–22nd January 2020 | Mathematical model | 6.49 |
| Liu et al. (2020) | China and overseas | 23rd January 2020 | Statistical exponential Growth using Poisson regression | 2.90 |
| Liu, Hu, et al. (2020) | China and overseas | 23rd January 2020 | Statistical maximum likelihood estimation | 2.92 |
| Read, Bridgen, Cummings, Ho, and Jewell (2020) | China | 1st–22nd January 2020 | Mathematical transmission model | 3.11 |
| Majumder & Mandl, 2020 | Wuhan | 8th December 2019–26th January 2020 | Mathematical Incidence Decay and Exponential Adjustment (IDEA) model | 2.0–3.1 ~2.55 |
| WHO | China | 18th January 2020 | The mathematical model including compartments Susceptible-Exposed-Infectious-Recovered-Death-Cumulative (SEIRD) | 1.4–2.5 ~1.95 |
| Cao et al. (2020) | China | 23rd January 2020 | – | 4.08 |
| Zhao et al. (2020) | China | 10th–24th January 2020 | Statistical exponential growth model corresponding to an eightfold increase in the reporting rate | 2.24 |
| Zhao et al. (2020) | China | 10th–24th January 2020 | Statistical exponential growth model corresponding to a twofold increase in the reporting rate | 3.58 |
| Imai et al. (2020) | Wuhan | 18th January 2020 | Mathematical model, computational modeling of potential epidemic trajectories | 1.5–3.5 ~2.5 |
| Riou and Althaus (2020) | China and overseas | 18th January 2020 | Stochastic simulations of early outbreak trajectories | 2.2 |
| Tang et al. (2020) | China | 22nd January 2020 | Mathematical SEIR-type epidemiological model | 6.47 |
| Li et al. (2020) | China | 22nd January 2020 | The statistical exponential growth model | 2.2 |
| Average R0 | | | | 3.28 |
$R_0$ value of 3.28 interprets that number of infected cases to double within 2 days. These studies consider the early period of December 2019 to January 2020, focusing mainly on cases in China (Sanche et al., 2020).

Real-time $R_0$ is defined by $R_t$ and the following figure shows $R_t$ for different US states. It is calculated on the go based on current reported new active cases and death numbers. The following figure shows a comparison of $R_t$ for different US states at an interval of 3 months which is May 2020 vs August 2020. We see $R_t$ has dropped for few states while it has increased for a few, this can be attributed to the implementation and relaxation of NPIs in these states. Fig. 5—$R_t$ for different US states shows contrast in $R_t$ for US states with a difference of 3 months. We can see the number of states which had relatively lower $R_t$ had bumped up, while states with higher $R_t$ have stabilized.

### 2.2 Incubation period and symptoms

Another aspect to look at when considering the spread of COVID-19 is the incubation period. The incubation period is the time between exposure to infection and symptom onset (Lauer et al., 2020). The estimated median
incubation period for COVID-19 is 5.1 days with a confidence interval of 95% in the range of 4.5–5.8 days. The majority of cases are 97.5% of cases will show symptoms within 11.5 days. Only 1% of cases will show symptom onset after 14 days. This research is based on confirmed COVID-19 cases reported in the period from 4th January 2020 to 24th February 2020. These results are also confirmed by the Centers for Disease Control and Prevention (CDC) (Management of patients with confirmed 2019-nCoV | CDC, n.d.) and also confirmed that the incubation period is extending to 14 days. From this, we can infer that a person identified as a confirmed COVID-19 patient would have contacted with infection in the last 14 days.

Based on Huang et al. (2020), Chen et al. (2020), and Pan et al. (2020), symptoms are shown by a person contracted with COVID-19 infection differ from person to person, but most people will show the following symptoms, majority reporting fever, cough, fatigue, and anorexia refer Fig. 6—Symptoms shown by COVID-19 cases. Apart from the given symptoms, few people have reported signs and symptoms with lower respiratory tract and gastrointestinal such as diarrhea and nausea before showing up with fever.

![Symptoms shown by COVID-19 cases](Management of patients with confirmed 2019-nCoV | CDC, n.d.)
In a study on 44,672 confirmed cases from the Chinese Center for Disease Control and Prevention performed by (Wu & McGoogan, 2020), the severity of COVID-19 is defined in three categories: mild, severe, and critical. Fig. 7—COVID-19 Illness Severity defines a split of confirmed COVID-19 cases among categories mild, severe, and critical with defined symptoms of each category.

2.3 Vaccine trials and availability

With the alarming rate at which COVID-19 is spreading, the World Health Organization (WHO) on 11th March 2020 declared COVID-19 as pandemic (WHO Timeline—COVID-19, n.d.). This also created tremendous pressure worldwide to get a vaccine for the same. Vaccines typically have a timeline of 18 months from the day of inception to be available for public use and have five testing process steps (Coronavirus Vaccine Tracker—The New York Times, n.d.). Fig. 8—Vaccine Testing Process shows different steps involved in vaccine development.

Different agencies throughout the world are rushing in developing vaccines at an increased pace with either combining Phase I and Phase II testing together or approving billions of dollars of funding before it is established that vaccine is efficient. Currently, there are over 140 vaccines in the development stage, with three in Phase III Efficacy Trials, refer to Fig. 9—Vaccine Development Status.
2.4 Non-pharmaceutical interventions (NPIs)

Even with the current rush on developing vaccines still, there is a good amount of time when the first vaccine will be available and till then other measures need to be put in place to curb the spread of this pandemic. The measures which help curb the spread of virus or infection apart from vaccine or medicine are known as Non-pharmaceutical Interventions (NPIs), which are also known as community mitigation strategies (Nonpharmaceutical Interventions (NPIs) | CDC, n.d.). NPIs define measures are personal, community, and environmental levels to curb the spread
of the virus, refer to Table 2 NPIs at different levels. Studies from the 1918 influenza pandemic in the United States show that the implementation of Non-pharmaceutical Interventions led to a reduction of mortality rate (Ferguson et al., 2006).

Applying one intervention in a silo does not create a lasting impact, a combination of multiple interventions is required to create a substantial impact (Ferguson et al., 2020). A combination of interventions can be applied using two broad strategies. 

1. **Mitigation**—this involves reducing the pace of the spread of pandemics resulting in reduced peak healthcare demand.

2. **Suppression**—this involves reducing new active case numbers and sustaining that going forward.

Comparing the two strategies, “Mitigation” will help in reducing peak healthcare demand. It will just slow down the spread of pandemics but won’t impact the number of cases. On the other hand, “Suppression,” which requires social distancing of the entire population, quarantine of cases and suspected cases, ban on community, and social gathering along with the closure of schools, universities, and workplaces, will reduce the new case numbers. For suppression to be successful, these interventions should be imposed till the vaccine is available else, whenever these are relaxed a spike in the number of cases will be observed. Since the vaccine development cycle is

| Personal NPI          | Community NPI                                                                 | Environmental NPI                                                                 |
|-----------------------|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Staying home when sick| Maintain social distancing                                                  | Routine cleaning or disinfecting of frequently touched objects or surfaces        |
| Avoid contact with sick people | Temporary closure of facilities which facilitate people are gathering such as schools, colleges, childcare center, sporting events, concerts, movie theaters, etc. |                                                                                  |
| Cover coughs and sneezes | Temporary closure of the workplace and encourage work from home wherever possible |                                                                                  |
| Washing hands        |                                                                             |                                                                                  |
| Cleaning/disinfecting frequently touched object or surface |                                                                             |                                                                                  |
| Routine cleaning or disinfecting of frequently touched objects or surfaces | |                                                                                  |

Table 2 NPIs at different levels.
18 months, imposing these interventions for such a long duration will come at social and economic costs. In place of implementing interventions for a long period, another way is to implement a combination of interventions in rotation for a small burst of the period, say 3 months, the author has researched implementing such rotation based small duration combination of interventions will result in reducing deaths by half and peak healthcare demand by two-thirds.

2.5 Disruptive technologies

Though we have identified NPI as an interim solution to curb the spread of COVID-19, data is still needed to support where NPIs need to be implemented. Required data is primarily to aid reasoning on whether the target population is likely to have COVID-19, this can be determined based on their social interaction or medical stats. Social interaction provides insights whether they have ever encountered anyone who is carrying COVID-19 or has been in contact with one, whereas medical stats will help in determining how to treat in case of identified positive for COVID-19. Listed below are identified technologies that can effectively implement NPIs (Abdel-Basset, Chang, & Nabeeh, 2021).

2.5.1 Artificial intelligence (AI) and machine learning (ML)

AI and ML are some of the most advanced technologies that are known to us. In these technologies, the machine learns based on existing observation and applies the same to the unknown to derive insights or inferences. These can be used in collaboration with robots for contactless, quick assessment, which is accurate and time-saving. Based on behavioral data, it can also segment users to the susceptibility of getting infected with COVID-19.

2.5.2 Industry 4.0 and internet of medical things (IoMT)

Internet of Things (IoT) is the basis of Industry 4.0 (Abdel-Basset, Nabeeh, El-Ghareeb, & Aboelfetouh, 2020). This brings the advent of medical cyber-physical systems to improve the quality and efficiency of medical services. 360-degree monitoring of the patient through multiple biosensors and applications helps in gaining meaningful insights in a time-efficient manner (Alloghani et al., 2018). This can also be leveraged for the early identification of COVID-19 and treatment of the same in novice state.
2.5.3 Virtual reality (VR), drones and autonomous robots

VR is another advanced technology that simulates a virtual environment, and this helps in creating refreshed, desired experience for the user. This helps psychologically in the current stay-at-home orders (Manto et al., 2020). Drones and autonomous robots are emerging with a vast variety of use cases in healthcare, one of the use cases is contactless high precision telesurgeries (Gupta, Shukla, & Tanwar, 2020; Howe & Matsuoka, 1999). Drones are also used to serve in remote areas for quick delivery of supplies and medical aid. 24 × 7 Patient observation and surveillance is another key benefit of drones in healthcare.

2.6 Contact tracing

Contact tracing is another important aspect to be taken into account to curb the spread of COVID-19. Any confirmed case is an active carrier of the virus since contact with the virus. Hence it becomes important to trace all the people with whom the confirmed case has been contacted in the last 14 days since the maximum incubation period is deemed 14 days for COVID-19 (Hellewell et al., 2020). For contact tracing and isolation between cases to be effective, the time lag between symptom onset and case isolation should be minimum, along with isolation of people who came in contact with a person showing symptoms. In the case of COVID-19 number of cases and contacts to be traced is huge and also, the availability of testing is not good. So, the feasibility of contact tracing and isolation of cases becomes questionable. In a study that was carried out in Taiwan, it was found that contact tracing and isolating symptomatic patients will only be effective if implemented together with other NPIs such as social distancing (Cheng et al., 2020).

3 Proposed approach

3.1 Dataset description

In this research, we are going to use four datasets providing information about confirmed cases, mobility factors, population, and daily reproduction numbers for each state. For this research, we are going to consider the date range from 8th March 2020 to 30th September 2020 for the country US at the state level. We are considering the given timeframe of data as we have harmonized data available for all US states during this period.
Table 3 Confirmed COVID-19 Cases Dataset Overview lists details around the “Confirmed COVID-19 Cases” dataset.

Table 4 COVID-19 Community Mobility Report Overview lists details around the “COVID-19 Community Mobility Report” dataset.

Table 5 US population—States Overview lists details around the “US population—States” dataset.

### Table 3 Confirmed COVID-19 cases dataset overview.

**Confirmed COVID-19 cases**

| Description | Contains data about confirmed cases daily for different countries. For the United States we have data available at the state and county level. Dataset is in pivot format with dates as columns. |
| Key features | • Date | • State |
| Key features | • Country | • Confirmed cases |

### Table 4 COVID-19 community mobility report overview.

**COVID-19 community mobility report**

| Description | Contains data about confirmed cases daily for different countries. For the United States we have data available at the state and county level. |
| Key features | • Date | • Country |
| Key features | • State | • Retail_and_recreation_percent_change_from_baseline |
| Key features | • Grocery_and_pharmacy_percent_change_from_baseline | • Parks_percent_change_from_baseline |
| Key features | • Transit_stations_percent_change_from_baseline | • Workplaces_percent_change_from_baseline |
| Key features | • Residential_percent_change_from_baseline | |

### Table 5 US population—States overview.

**US population—States**

| Description | Contains percentage change in mobility factors from baseline for different countries. For the United States, data is available at the state level. |
| Key features | • State |
| Key features | • Population |
| Key features | • Population density |
Table 6 Daily $R_0$ value for US states lists details around the “Daily $R_0$ value—US states” dataset.

All of these datasets are available in the public domain and are free to use. All datasets except the population dataset get updated daily.

### 3.1.1 Smoothing weekend effect

The “Confirmed COVID-19 cases” dataset captures daily new confirmed cases of COVID-19. We notice a pattern where the number of confirmed cases reduces drastically on weekends (Saturday and Sunday) and gets a spike on Monday. This is due to limited testing on weekends and cascading weekend testing to Monday. To harmonize the number of new confirmed cases over the week to eradicate such behavior, we will average out cases from the past 2 days to the future 2 days as shown in Fig. 10—Data transformation of Confirmed COVID-19 cases. Two days past and future are considered to level out any effect of the weekend, which is Saturday and Sunday.

| Description | Contains R$_0$ values for each state daily |
|-------------|-------------------------------------------|
| Key features | • Date |
|             | • Region (State) |
|             | • Mean R$_0$ |

Table 6 Daily $R_0$ value for US states.

Fig. 10 Data transformation of confirmed COVID-19 cases.
3.2 Feature engineering

During our literature review, we found out that the incubation period COVID-19 virus extends up to 14 days (Management of patients with confirmed 2019-nCoV | CDC, n.d.) from which we can infer a person identified as COVID-19 infected today would have contacted with the virus in past 14 days. Hence, we can consider moving the sum or average of features in the last 14 days as new features to predict confirmed cases today.

3.3 Correlation: Confirmed_cases vs other features

Above Fig. 11—Correlation Confirmed_Cases vs other features shows the correlation of Confirmed_Cases with other features. We notice a relatively good correlation with Cases_past_14_days_mean, Cases_past_14_days_sum, and population which attributes to the fact that COVID-19 spreads from person to person, so an infected person is a carrier of the virus. Hence existing infected cases are directly proportional to new confirmed cases. One of the interesting observations is it is negatively correlated with population density (pop_density) which infers population density does not have any impact on new confirmed cases.

In terms of correlation with mobility features, staying at home is helping curb new confirmed cases, while mobility in retail and recreation areas and, workplaces are helping COVID-19 confirmed cases to bump.

Fig. 11 Correlation confirmed_cases vs other features.
Basic Reproduction Rate ($R_0$) seems to have a slight negative correlation implying to be an independent feature and does not help in defining new confirmed COVID-19 cases.

4 Experiment and results

4.1 Model building

This is a regression problem for which the following models are considered for evaluation.

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Light Gradient Boosting
- Extreme Gradient Boosting
- AdaBoost

Metrics used for evaluation are $R^2$ and RMSE; these regressor models performed as defined in Table 7 Model Performance without tuning.

Based on the performance of these models, we shortlist models that have $R^2 > 0.99$ for further tuning. Shortlisted models based on these criteria are

- Random Forest
- Light Gradient Boosting
- Extreme Gradient Boosting

Comparing the performance of tuned models against base models.

Table 8 Base Models vs Tuned Models shows that base models without any hyperparameter tuning perform better than tuned models. Base models have high $R^2$ and low RMSE than tuned models.

Evaluating each of these base models in detail.

| Model                        | RMSE  | $R^2$  |
|------------------------------|-------|--------|
| Linear Regression            | 0.0958| 0.486  |
| Decision Tree                | 0.0127| 0.9895 |
| Random Forest                | 0.0077| 0.9961 |
| Gradient Boosting            | 0.0253| 0.9641 |
| Light Gradient Boosting      | 0.0088| 0.9951 |
| Extreme Gradient Boosting    | 0.0085| 0.9957 |
| AdaBoost                     | 0.0655| 0.7584 |
Table 8  Base models vs tuned models.

| Model                | R2       | RMSE  | Model                | R2       | RMSE  |
|----------------------|----------|-------|----------------------|----------|-------|
| Random Forest        | 0.9961   | 0.0077| Random Forest        | 0.9955   | 0.0088|
| Light Gradient       | 0.9951   | 0.0088| Light Gradient       | 0.9947   | 0.0093|
| Boosting             |          |       | Boosting             |          |       |
| Extreme Gradient     | 0.9957   | 0.0085| Extreme Gradient     | 0.927    | 0.0361|
| Boosting             |          |       | Boosting             |          |       |

Table 9  Random forest—Hyperparameters.

| Hyperparameter             | Value       | Hyperparameter             | Value       |
|---------------------------|-------------|---------------------------|-------------|
| bootstrap                 | TRUE        | min_samples_leaf          | 1           |
| ccp_alpha                 | 0           | min_samples_split         | 2           |
| criterion                 | “mse”       | min_weight_fraction_leaf  | 0           |
| max_depth                 | None        | n_estimators              | 100         |
| max_features              | “auto”      | n_jobs                    | -1          |
| max_leaf_nodes            | None        | oob_score                 | FALSE       |
| max_samples               | None        | random_state              | 5992        |
| min_impurity_decrease     | 0           | verbose                   | 0           |
| min_impurity_split        | None        | warm_start                | FALSE       |

4.1.1 Random forest

Hyperparameters

Table 9 Random forest—Hyperparameters states hyperparameters used by Random Forest Regressor.

Residuals and prediction error plot

Above Fig. 12—Random Forest—Residuals and Prediction Error Plot shows residuals and prediction error plots for Random Forest regressor. Residuals for train and test data exhibit a similar trend within the range of ±0.05. Prediction errors are in line with the best fit line with a very slight deviation having an R2 of 0.994.

Learning curve plot

Fig. 13—Random Forest—Learning Curve Plot shows the learning curve for Random Forest regressor. We notice that till 4000 instances learning algorithm keeps improving post which it saturates.
4.1.2 Light gradient boosting

Hyperparameters

Table 10 Light Gradient Boosting—Hyperparameters show the hyperparameters of the model.

Residuals and prediction error plot

Fig. 14—Light Gradient Boosting—Residuals and Prediction Error Plot shows that residuals for train and test datasets exhibit a similar trend within the range of ±0.1. Prediction Errors follow the best fit line with R² of 0.994.

Learning curve plot

Fig. 15—Light Gradient Boosting—Learning Curve Plot shows a rise in test data score up to 5700 instances post which falls steeply.
Table 10 Light gradient boosting—Hyperparameters.

| Hyperparameter          | Value          | Hyperparameter          | Value  |
|-------------------------|----------------|-------------------------|--------|
| boosting_type           | "gbdt"         | n_jobs                  | -1     |
| class_weight            | None           | num_leaves              | 31     |
| colsample_bytree        | 1              | objective               | None   |
| importance_type         | "split"       | random_state            | 5992   |
| learning_rate           | 0.1            | reg_alpha               | 0      |
| max_depth               | -1             | reg_lambda              | 0      |
| min_child_samples       | 20             | silent                  | TRUE   |
| min_child_weight        | 0.001          | subsample               | 1      |
| min_split_gain          | 0              | subsample_for_bin       | 200000 |
| n_estimators            | 100            | subsample_freq          | 0      |

Fig. 14 Light gradient boosting—Residuals and prediction error plot.

Fig. 15 Light gradient boosting—learning curve plot.
4.1.3 Extreme gradient boosting

Hyperparameters

Table 11 Extreme Gradient Boosting—Hyperparameters show hyperparameters used by the model.

Residuals and prediction error plot

From Fig. 16—Extreme Gradient Boosting—Residuals and Prediction Error Plot, we notice that residuals are limited between the range of ±0.05 for both the train and test dataset. Prediction error follows the best fit line with R² of 0.993.

Table 11 Extreme gradient boosting—hyperparameters.

| Hyperparameter              | Value       | Hyperparameter              | Value       |
|-----------------------------|-------------|-----------------------------|-------------|
| base_score                  | 0.5         | monotone_constraints        | “[]”        |
| booster                     | “gbtree”    | n_estimators               | 100         |
| colsample_bylevel           | 1           | n_jobs                      | -1          |
| colsample_bynode             | 1           | num_parallel_tree           | 1           |
| colsample_bytree             | 1           | objective                   | “reg:squarederror” |
| gamma                       | 0           | random_state                | 5992        |
| gpu_id                      | -1          | reg_alpha                   | 0           |
| importance_type             | “gain”      | reg_lambda                  | 1           |
| interaction_constraints     | “”          | scale_pos_weight            | 1           |
| learning_rate               | 0.300000012 | subsample                   | 1           |
| max_delta_step              | 0           | tree_method                 | “exact”     |
| max_depth                   | 6           | validate_parameters         | 1           |
| min_child_weight            | 1           | verbosity                   | 0           |
| missing                     | nan         |                             |             |

Fig. 16 Extreme gradient boosting—residuals and prediction error plot.
Learning curve plot
From Fig. 17—Extreme Gradient Boosting—Learning Curve Plot, it can be deduced that the learning curve increases steeply up to 4000 instances and then saturates.

4.2 Model interpretability

4.2.1 Feature importance

Feature importance is defined by the learning algorithm and is based on the importance of a feature in predicting the target variable.

Random forest
Based on Fig. 18—Feature Importance—Random Forest, we can infer new confirmed COVID-19 cases have maximum impact from confirmed cases in the past 14 days, followed by population density and population. In mobility parameters, transit stations were the highest contributor to new confirmed cases. The basic reproduction rate has a very low impact as compared to other features.

Light gradient boosting
Based on Fig. 19—Feature Importance—Light Gradient Boosting, we can infer new confirmed COVID-19 cases have maximum impact from confirmed cases in the past 14 days, followed by population density and population. In mobility parameters, residential, transit stations, and parks have
good and nearly the same impact, while retail and recreation have the least effect. The basic reproduction rate also has a considerable effect on new confirmed cases.

**Extreme gradient boosting**

Based on **Fig. 20**—Feature Importance—Extreme Gradient Boosting, we can infer new confirmed COVID–19 cases have maximum impact from confirmed cases in the past 14 days, followed by population density and population. In mobility parameters, transit stations have a maximum. The basic reproduction rate has a low effect on new confirmed cases.

---

**Fig. 18** Feature importance—Random forest.

**Fig. 19** Feature importance—Light gradient boosting.
4.2.2 SHAP interpretability

**Random forest**

Based on Fig. 21—SHAP Interpretability—Random Forest, we can infer new confirmed COVID-19 cases have maximum impact from confirmed cases in the past 14 days, population density, and population. In mobility parameters, grocery and pharmacy had the most effect on new confirmed cases. The basic reproduction rate has a low impact as compared to other features.
Light gradient boosting
Based on Fig. 22—SHAP Interpretability—Light Gradient Boosting, we can infer new confirmed COVID-19 cases have maximum impact from confirmed cases in the past 14 days followed by population density and population. In mobility parameters, residential and transit stations have a good impact. The basic reproduction rate also has a considerable effect.

Extreme gradient boosting
Based on Fig. 23—SHAP Interpretability—Extreme Gradient Boosting, we can infer new confirmed COVID-19 cases have maximum impact from confirmed cases in the past 14 days followed by population density and population. In mobility parameters, residential and transit stations have a good impact. The basic reproduction rate also has a considerable effect.

4.2.3 LIME interpretability
Random forest
Based on Fig. 24—LIME Interpretability—Random Forest, we can infer a positive impact on new confirmed cases is from the population, followed by mobility parameters—retail and recreation, and transit stations. Cases from the past 14 days have a positive impact along with the basic reproduction rate. The remaining features have a negative effect on new confirmed cases.

Fig. 22 SHAP interpretability—Light gradient boosting.
Light gradient boosting

Based on Fig. 25—LIME Interpretability—Light Gradient Boosting, we can infer a positive effect on new confirmed cases is from population followed by mobility parameters—workplaces, retail and recreation, and transit stations. Cases from the past 14 days have a positive impact along with the basic reproduction rate. The remaining features have a negative effect on new confirmed cases.

Extreme gradient boosting

Based on Fig. 26—LIME Interpretability—Extreme Gradient Boosting, we can infer a positive effect on new confirmed cases is from population followed by mobility parameters—retail and recreation, and transit stations.
Cases from the past 14 days have a positive impact along with the basic reproduction rate. The remaining features have a negative effect on new confirmed cases.

4.3 Model evaluation

This study is about finding the effect of different features on new confirmed cases, which is done by creating statistical regression models predicting new confirmed cases and comparing the same with actual numbers. Metrics used for model evaluation are Root Mean Square Error (RMSE) and the Coefficient of Determination ($R^2$).

The following regression statistical regression models are considered for evaluation in this study:

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Light Gradient Boosting
- Extreme Gradient Boosting
- Adaboost
Fig. 27—Model performance—Base Models, we notice high-performing models are Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting, which have similar performance with $R^2$ in the range of $0.9951-0.9961$ and RMSE in the range of $0.0077-0.0088$. Followed by Decision Tree with $R^2$ as $0.9895$ and RMSE as $0.0127$, and Gradient Boosting with $R^2$ as $0.9641$ and RMSE as $0.0253$. The model that performed the worst is Linear Regression with $R^2$ as $0.486$ and RMSE as $0.0958$, and AdaBoost with $R^2$ as $0.7584$ and RMSE as $0.0655$.

Based on this, further tuning modes are selected with a cut-off on $R^2$ as $0.99$ and shortlisted models are Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting. These models even have the lowest RMSE.

Fig. 28—Model Performance—Base vs Tuned Models compares the performance of models with and without tuning. We notice a dip in performance of all the models post tuning and a major dip is shown by Extreme Gradient Boosting with an approximately 7 points dip in $R^2$. Random Forest and Light Gradient Boosting after tuning show a dip in performance which is comparable and is not drastic. Based on this observation,
Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting are considered without tuning for model interpretation. In terms of RMSE, all three models have similar performance with a slight deviation. The model that performed best in terms of RMSE and $R^2$ is Random Forest, with RMSE as 0.0077 and $R^2$ as 0.9961.

4.4 Model interpretability

As part of model interpretability methods used are feature importance from models, SHAP interpretability, and LIME interpretability. Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting are all interpreted on given interpretability methods. The importance of a feature in explaining target variable new confirmed cases is explained. LIME method predicted values that are in the range of 49–62% of actual values, hence interpretation of LIME could be overridden by other interpretability methods, which make a strong and better case.

Model Interpretability for Random Forest statistical model is shown in Fig. 28—Model performance—Base vs tuned models.
Fig. 29 Random forest—Model interpretability.
in Fig. 30—Light Gradient Boosting—Model Interpretability and Extreme Gradient Boosting in Fig. 31—Extreme Gradient Boosting—Model Interpretability. Model Interpretability against each feature is detailed out below.

### 4.4.1 Confirmed cases in the last 14 days
From all three statistical models’ interpretability, it can be observed that total confirmed cases in the past 14 days have maximum effect on new confirmed cases that is it contributes to an increasing number of new confirmed cases, which is also in agreement with the fact that COVID-19 spreads from person-to-person and can be spread only from existing confirmed cases. Mean confirmed cases in the past 14 days feature has a high correlation with total confirmed cases in the past 14 days; hence it received either similar importance as total confirmed cases or was dropped due to fact of correlation.

### 4.4.2 Population and population density
It can be observed from all three models’ interpretability that the population also contributes to an increase in new confirmed cases with a good magnitude which infers that highly populated locations are prone to have a steep rise in new confirmed cases. For population density, feature importance and SHAP interpretability for all three models depict a high positive impact on new confirmed cases, but LIME interpretability for all three models depicts a slight negative impact that contradicts feature importance and SHAP interpretability. It can also be observed LIME prediction is in the range of 49–62%. Hence we can override LIME observation with feature importance and SHAP interpretability. Based on this, highly populated locations are also prone to a steep increase in new confirmed cases, which is again in alignment with person-to-person transmission of COVID-19.

### 4.4.3 Basic reproduction rate
The basic reproduction rate positively impacts new confirmed cases from the model interpretability of all three statistical models. So an increase in basic reproduction rate can be deduced as rising in new confirmed cases. This is also in alignment with the definition of basic reproduction rate, according to which it defines the rate at which pandemic is expanding. This feature can be controlled only by curbing the spread of COVID-19.
Fig. 30 Light gradient boosting—Model interpretability.
Fig. 31 Extreme gradient boosting—Model interpretability.
4.4.4 Global mobility features

Based on model interpretability from all three models, we have created a formula based on a weighted average to define feature importance.

Assign the value to each feature based on the impact it draws on new confirmed cases ranging from 1 to 6, with 1 being least impact and 6 with maximum impact. Feature value 0 is assigned either feature impact on new confirmed cases is negative, or the interpretability method does not assign any value that drops the feature from calculations. Weights are assigned based on statistical algorithms and interpretability used. Since all three statistical models have similar performance, so weight is going to be 1. For weights about interpretability, feature importance is derived from statistical models; hence it will be given weight as 1, and SHAP is also given weightage as 1. LIME, on the other hand, has a prediction in the range of 49–62% hence is given a weight of 0.5.

Equation is

\[
\text{Net feature effect} = \frac{\sum_{i=1}^{m} M_k \left( \sum_{j=1}^{n} F_{ij} \times V \right)}{m}
\]

Where,

\( M = \) Model which in this case is Random Forest, Light Gradient Boosting, Extreme Gradient Boosting.

\( F = \) Model Interpretability method, which is Feature Importance, SHAP, LIME.

\( m = \) number of models.

\( n = \) number of model interpretability methods.

\( V = \) Feature value in the range of features.

From Table 12 Global Mobility Features Importance, it can be inferred that the mobility of people in transit stations is contributing to the new confirmed cases more than all other mobility features. Grocery, pharmacy, and residential places stand second in maximum contribution to new confirmed cases, followed by retail and recreation places and parks. Workplaces stand last, which is since maximum corporate companies have enabled remote working as part of an effort to curb the spread of COVID-19.

Transit stations that top the list for contributing to new confirmed cases need some serious considerations to curb the spread of COVID-19. During onboarding and deboarding times, social distancing has been on the toss, which might be one of the reasons for the spread. Also, the population
Table 12 Global mobility features importance.

| Global mobility feature | Random forest | | | | Light gradient boosting | | | | | | Extreme gradient boosting | | | | | | | | | | Net | feature | effect |
|-------------------------|---------------|---|---|---|--------------------------|---|---|---|---|--------------------------|---|---|---|---|--------------------------|---|---|---|---|---|
| Retail and recreation   | 2             | 2 | 6 | 1 | 4 | 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Grocery and pharmacy    | 5             | 6 | 0 | 3 | 4 | 0 | 5 | 4 | 0 | 3 | 3 | 3 | 3 | 3 |
| Transit stations        | 6             | 4 | 5 | 5 | 5 | 6 | 6 | 5 | 6 | 4 | 4 | 4 | 4 | 4 |
| Parks                   | 3             | 3 | 0 | 4 | 3 | 0 | 2 | 3 | 0 | 2 | 2 | 2 | 2 | 2 |
| Workplaces              | 0             | 1 | 0 | 2 | 0 | 5 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| Residential             | 4             | 5 | 0 | 6 | 6 | 0 | 4 | 6 | 0 | 3 | 3 | 3 | 3 | 3 |
dependent on transit stations is relatively high. Grocery and pharmacy is a place where the population approach for daily essentials. During peak time, traffic to such places increases interaction between populations, making them prone to COVID-19. Lately, a lot of family get-togethers are observed, which is again another place of interaction. Parks contributing is significantly less due to the self-isolation meets that can be done in parks due to the massive size of parks.

5 Conclusion and future work

5.1 Discussion and conclusion

This study deals with an ongoing pandemic known as COVID-19 and aims to find non-pharmaceutical interventions that can be implemented in the absence of a vaccine to curb the spread of the virus. An extensive literature review compared to COVID-19 pandemic with previous pandemics and enlightened with different facts about this virus: it spreads from person to person, has an incubation period of up to 14 days, and has a high reproduction rate. The incubation period of COVID-19 is one of the strange aspects to notice, which means a person might not show symptoms of COVID-19 up to 14 days of contraction and in rare cases, this has extended up to 30 days as well. This thing makes controlling the spread of this virus a daunting task. Hence it needs a different approach as compared to previous pandemics to control.

The literature review also discusses vaccine trials, it is found that a typical vaccine has a lifecycle of 18 months with a five-step approach. With the rush to develop a vaccine, steps for the vaccine are accelerated but can’t be skipped and with that, the vaccine will take a good amount of time to be available. Post availability of vaccine another set of hurdles of mass production and distribution brings in the next set of challenges. With all of this, curbing COVID-19 in the absence of vaccines becomes a prime objective and steps to keep social and economic balance becomes important. These steps to curb the spread of the virus in the absence of a vaccine are known as Non-pharmaceutical Interventions (NPIs) and the goal of this study is to find NPIs that need to be implemented to curb the spread of COVID-19.

NPIs can be implemented at either personal level, community level, or environmental level and can have the effect of either mitigation (reducing the pace of the spread of pandemic) or suppression (reducing new confirmed cases and sustaining the same). Personal NPIs evolve from self-awareness and responsibility, it includes self-quarantine in case of sick, avoids interaction
with people, washing hands, cleaning, or disinfecting frequently touched objects. Community NPIs are general closure or restriction on traffic in public places to reduce population interaction, implement social distancing, and restrict mass gathering. Implementing NPIs curbs the spread of the virus but also brings a good impact on the social and economic conditions of society. Hence NPIs need to be implemented in a balanced way that has maximum effect on curbing the spread of the virus but the least social and economic effect.

This study considers data from the following four different sources.

1. Confirmed COVID-19 cases
   Provides data about new confirmed cases on a daily basis for each US state.

2. COVID-19 community mobility report
   Provides data on the mobility of the population in different places for each US state.

3. US population—States
   Provides data about population and population density for each state in the US.

4. Daily R₀ value—US states
   Provides daily reproduction rate of COVID-19 for each US state.

It further considers only specific feature from each of these data sources to created one integrated dataset for analysis which includes features on daily new confirmed cases, population, population density, daily basic reproduction rate, and global mobility factors for retail and recreation, grocery, and pharmacy, transit stations, parks, workplaces, and residential places. The data is at state granularity on a daily basis and is considered for dates 8th March 2020 to 30th September 2020, both days inclusive. The integrated dataset is taken through multiple pre-processing steps, one the important being smoothening of confirmed cases over ±2 days to eradicate the weekend effect during the weekend. It was observed low to no cases while Monday used to have an abnormal peak, which was since limited testing was available on the weekend. Outliers were not removed as the study wants to consider all such scenarios. Feature engineering involved considering all features in the past 14 days since the probability of a person found to confirm the case today might have contracted a virus in the last 14 days based on the incubation period.

This is a regression problem aiming to forecast the anticipated number of new COVID-19 cases on a given day in the United States. Based on which relation of a feature on target variable is deduced to define the causal
relationship. For this study, stratified K-fold cross-validation with an ensemble of the following models is evaluated.

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Light Gradient Boosting
- Extreme Gradient Boosting
- Adaboost

These models were evaluated based on metrics Root Mean Square Error (RMSE) and Coefficient of Determination ($R^2$). It was observed Random Forest, Light Gradient Boosting, and Extreme Gradient Boosting performed at par with similar performance while tuned versions of these models had relatively similar performance on the lower side except for Extreme Gradient Boosting, where it dropped significantly. The best performing models on these metrics are Random Forest with RMSE as 0.0077 and $R^2$ as 0.9961 followed by Extreme Gradient Boosting with RMSE as 0.0085 and $R^2$ as 0.9957 and Light Gradient Boosting with RMSE as 0.0088 and $R^2$ as 0.9951.

Further model interpretability explained the effect of each feature on new confirmed cases. Three different methods were used to understand the effect of features on new confirmed cases that were feature importance derived from the model, SHAP interpretability, and LIME interpretability. It was observed that the LIME method predicted values in the range of 49–62%; hence interpretations made by LIME were overridden by other interpretability methods, which made a better and strong case. Based on model interpretability following are the observations.

- **Confirmed cases in the past 14 days**

  It was found that confirmed cases in the past 14 days contribute maximum to new confirmed cases, which are also in alignment with COVID-19 spreads from person to person. This also infers that to curb the spread of the virus, quarantine of confirmed cases needs to follow until recovery strictly. Also contact tracing should be done to tag possible new confirmed cases and a strict quarantine for them as well needs to be ordered until the incubation period is over, along with a negative COVID-19 test.

- **Population and population density**

  It was found that both population and population density contributes to the spread of COVID-19. Hence locations with a high
population and population density are more prone to have a steep rise in new confirmed cases. Social distancing should be followed along with the implementation of personal, community, and environmental NPIs.

- **Basic reproduction rate**
  
The basic reproduction rate is derived from the pace of the spread of the virus, it showed a positive impact on new confirmed cases. An increase in the basic reproduction rate can be attributed to the high forecast in new confirmed cases.

- **Global mobility features**
  
  Population mobility in retail and recreation, grocery and pharmacy, transit stations, parks, workplaces, and residential places is considered as part of this study. It was found that mobility in transit stations contributes maximum to new confirmed cases followed by grocery and pharmacy, and residential. Retail and recreation, parks, and workplaces have a very low contribution to new confirmed cases. Hence community and environmental NPIs need to be implemented in transit stations. Frequent cleaning and disinfection, strict social distancing, and restriction of the cluster of people. For groceries and pharmacy delivery and pickup options through the drive-through need to be encouraged. The number of people in-store should be restricted to a basic minimum to follow social distancing. With the oncoming of the holiday, season family gets together are on the rise, for the same personal NPIs should be encouraged to maintain the safety of everyone. Workplaces are already helping curb the spread by pushing remote working, retail, and recreation to use drive-through or open with limited capacity. Parks should also restrict and create small bubbles to avoid mass interactions. Table 13 Conclusion—NPIs identified, summarises NPIs that need to be implemented as part of this study to supress and mitigate the spread of COVID-19.

5.2 Contribution to knowledge

- **Strategy to find NPIs to mitigate and suppress COVID-19**
  
  This study defines the process of converting data insights into action-able items of implementing which NPI will yield maximum benefit in terms of curbing the spread of COVID-19. The process is based on data available in the public domain and then transform the data and engineer new features based on domain knowledge and factors that define
COVID-19. Once high influential factors are identified, corresponding NPIs can be implemented to curb the spread of the virus. This strategy helps mitigate the spread of the virus to reduce peak healthcare demand and suppress the virus until the vaccine is available.

- **Strategy to find places that are more susceptible to influence the spread of COVID-19**

  Based on global mobility factors, it defines the strategy of finding the effect of population mobility in a specific place to new confirmed cases. It was also observed that places with a high population and population density are more prone to a steep rise in new confirmed cases. Based on these two observations, strategy or NPI to restrict population mobility and interaction in places that are more prone to contribute to new confirmed cases can be identified.

- **Weighted average to find net feature effectiveness-based interpretability from multiple methods**

  This study used three different model interpretability methods to find feature importance to target variables and then defined a method of inferring net feature importance based on the weighted average.

| Situation/place                        | NPI                                                                 |
|----------------------------------------|----------------------------------------------------------------------|
| Existing confirmed cases               | Stringent quarantine until recovery with a negative test result     |
| Person with possible exposure to the virus | Quarantine up to 14 days with negative COVID-19 test               |
| Places with high or dense population   | Personal and community NPIs—includes self-awareness, social distancing, restriction on mass gathering |
| Transit stations                       | Social distancing, restrict peak hour rush, restrict entry to only for important use |
| Grocery and pharmacy                   | Social distancing, restrict on maximum occupancy, promote drive through pickups or delivery |
| Residential                            | Personal NPIs, promote self-awareness                              |
| Retail and recreation                   | Personal and Community NPIs—includes self-awareness, social distancing, restriction on mass gathering |
| Parks                                  | Personal and Community NPIs—includes self-awareness, social distancing, restriction on mass gathering |
| Workplaces                             | Sustain remote working till vaccine is available                   |

Table 13 Conclusion—NPIs identified.
Weights were defined to statistical algorithms used based on performance and model interpretability methods based on confidence and relevance of the same. This strategy will help integrate interpretability received from multiple sources and define net feature importance.

### 5.3 Future recommendations

- **Use testing effectiveness as one of the features to define new confirmed cases**
  
  Testing is one of the key factors that define existing confirmed cases, and adequate testing will increase relevance while gaps in testing are going to create voids. Considering testing data is going to increase visibility on existing confirmed cases and deviation of the same from actual numbers. This will further narrow down NPIs that need to be implemented for effective control of COVID-19.

- **Define NPIs at the state level and compare them against country level**
  
  The current study uses data at the country level, take data one level deep is going defining a more personalized and focused strategy for a state taking into account state-focused limitations. The densely populated state has a different set of challenges as compared to the sparingly populated state. For example, the pace of the spread of COVID-19 in the densely populated state is high, whereas testing is a limitation in the sparingly populated state.

### Credit authorship contribution statement

Suvajit Mukhopadhyay: Supervision and review.

Victor Chang: Review and guideline.

### References

Abdel-Basset, M., Chang, V., & Nabeeh, N. A. (2021). An intelligent framework using disruptive technologies for COVID-19 analysis. *Technological Forecasting and Social Change, 163*, 120431.

Abdel-Basset, M., Nabeeh, N. A., El-Ghareeb, H. A., & Aboelfetouh, A. (2020). Utilising neutrosophic theory to solve transition difficulties of IoT-based enterprises. *Enterprise Information Systems, 14*, 1304–1324.

Alloghani, M., Al-jumeily, D., Hussain, A., Aljaaf, A. J., Mustafina, J., & Petrov, E. (2018). Healthcare services innovations based on the state of the art technology trend industry 4.0.2018. In *11th int. conf. dev. esystems eng* (pp. 64–70).

Cao, Z., Zhang, Q., Lu, X., Pfeiffer, D., Jia, Z., Song, H., et al. (2020). *Estimating the effective reproduction number of the 2019-nCoV in China*, medRxiv. Cold Spring Harbor Laboratory Press.
Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine, 27*, 1–4.

Liu, T., Hu, J., Xiao, J., He, G., Kang, M., Rong, Z., et al. (2020). Time-varying transmission dynamics of novel coronavirus pneumonia in China. *bioRxiv*. 2020.01.25.919787.

Management of patients with confirmed 2019-nCoV | CDC [WWW Document], n.d. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html (Accessed 27 June 2020).

Majumder, M., & Mandl, K. D. (2020). Early transmissibility assessment of a novel coronavirus in Wuhan, China. *SSRN Electron Journal*. https://dx.doi.org/10.2139/ssrn.3524675.

Manto, M., Dupre, N., Hadjivassiliou, M., Louis, E. D., Mitoma, H., Molinari, M., et al. (2020). Management of Patients with cerebellar Ataxia during the COVID-19 pandemic: Current concerns and future implications. *Cerebellum, 19*, 562–568.

Nonpharmaceutical Interventions (NPIs) | CDC [WWW Document], n.d. https://www.cdc.gov/nonpharmaceutical-interventions/index.html (Accessed 20 June 2020).

Moustakas, A. http://arxiv.org/abs/2006.00971.

Ong, S. W. X., Tan, Y. K., Chia, P. Y., Lee, T. H., Ng, O. T., Wong, M. S. Y., et al. (2020). Air, surface environmental, and personal protective equipment contamination by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from a symptomatic patient. *JAMA, the Journal of the American Medical Association, 323*(16), 1610–1612.

Pan, L., Mu, M., Yang, P., Sun, Y., Wang, R., Yan, J., et al. (2020). Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: A descriptive, cross-sectional, multicenter study. *The American Journal of Gastroenterology, 115*, 766–773.

Read, J. M., Bridgen, J. R. E., Cummings, D. A. T., Ho, A., & Jewell, C. P. (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *medRxiv*. 2020.01.23.20018549.

Riou, J., & Althaus, C. L. (2020). Pattern of early human-to-human transmission of Wuhan 2019-nCoV. *bioRxiv*. 2020.01.23.917351.

Rt: Effective Reproduction Number [WWW Document], n.d. https://rt.live/ (Accessed 16 August 2020).

Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N., & Ke, R. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases, 26*, 1470–1477.

Shen, M., Peng, Z., Xiao, Y., & Zhang, L. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. *bioRxiv*. 2019, 2020.01.23.916726.

Sun, J., He, W. T., Wang, L., Bai, A., Ji, X., Zhou, X., et al. (2020). COVID-19: Epidemiology, evolution, and cross-disciplinary perspectives. *Trends in Molecular Medicine, 26*, 483–495.

Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y., et al. (2020). Estimation of the transmission risk of 2019-nCoV and its implication for public health interventions. *SSRN Electron Journal*. https://doi.org/10.2139/ssrn.3525588.

van den Driessche, P., & Watmough, J. (2008). Further notes on the basic reproduction number. *Journal of Heredity, 159–178*. Springer Verlag.

WHO coronavirus disease (COVID-19) dashboard | WHO Coronavirus Disease (COVID-19) Dashboard[WWW Document], n.d. https://covid19.who.int/ (Accessed 29 August 2020).

WHO Timeline—COVID-19 [WWW Document], n.d. URL https://www.who.int/news-room/detail/27-04-2020-who-timeframe---covid-19 (Accessed 27 June 2020).

WHO | Middle East respiratory syndrome coronavirus (MERS-CoV) (Ed.). (2020). *WHO*. WHO | Summary of probable SARS cases with onset of illness from 1st November 2002 to 31st July 2003 (Ed.). (2015). *WHO*.
Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet*, 395, 689–697.

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. *JAMA*, 323, 1239.

Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., et al. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *92* (pp. 214–217). Elsevier.