2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00209

Pandemic spread prediction and healthcare preparedness through financial and mobility data

Nidhi Mulay*, Vikas Bishnoi†, Himanshi Charotia‡, Siddhartha Asthana§, Gaurav Dhama¶, and Ankur Arora‖
AI Garage, Mastercard
DLF Plaza Tower, DLF Phase 1, Sector 26A, Gurugram, Haryana 122002, India
Email: {nidhi.mulay*,vikas.bishnoi†,himanshi.charotia‡,siddhartha.asthana§,gaurav.dhama¶,ankur.arora‖}@mastercard.com

***Abstract*—The pandemics like Coronavirus disease 2019 (COVID-19) require Governments and health professionals to make time-sensitive, critical decisions about travel restrictions and resource allocations. This paper identifies various factors that affect the spread of the disease using transaction data and proposes a model to predict the degree of spread of the disease and thus the number of medical resources required in upcoming weeks. We perform a region-wise analysis of these factors to identify the control measures that affect the minimal set of population. Our model also helps in estimating the surges in clinical demand and identifying when the medical resources would be saturated. Using this estimate, we suggest the preventive as well as corrective measures to avoid critical situations.**

***Index Terms*—machine learning; visual analysis; COVID-19; social distancing; health; confirmed cases; regression; counterfactual examples**

I. Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).It was first identified in Wuhan, Hubei, China, during December 2019 and has since been spread worldwide resulting in a global pandemic. More than 21 million cases have been reported across 188 countries and territories as of 15 August 2020, resulting in more than 7,62,000 deaths.[1]

Authorities worldwide have responded by implementing travel restrictions, home quarantine, workplace hazard controls, and facility closures since pathogens can spread exponentially without these pandemic containment measures of social distancing. Due to these quarantine orders and closure of many industries, this outbreak is causing a major destabilising threat to the global economy. Hence, it has become vital to analyze the scenario at granular level and take decisions keeping the effects on the economy within manageable levels, and simultaneously flattening the epidemic curve.

In this paper, we aim to analyze the pattern in citizens' movement post the quarantine orders so that the officials can make efficient decisions of whether a particular region needs stricter rules, more facilities or whether the citizens of a region are responsible and aware enough to be exempted from the quarantine rules and regulations. We also present a machine learning model to predict the new confirmed cases in following week to notify the health workers about the medical resources that might be needed in future. A challenge to building a Machine Learning model is the lack of historical data. Many epidemiological models have been developed for policymakers, clinicians, and health practitioners but most of these use the historical data from other coronaviruses such as SARS and MERS while some use the publicly available datasets related to COVID cases and availability of medical resources. Pandemics are rare and have different characteristics so the data of other coronaviruses cannot be used for Covid-19[40]. Hence, we have created additional predictive features for our model by analyzing their effect on the spread of the disease. Along with some public datasets, we also use *Mastercard Transaction dataset* as it provides us information about the response of citizens to the pandemic control guidelines based on their pattern of purchasing. Analysis of this pattern gives insights about its impact on the spread of the disease. We mainly focus on the following tasks in order to determine the factors responsible for COVID spread in any county:

- We first identify the factors that might affect the spread of COVID and then create predictive variables based on these factors by merging datasets from different sources. We categorize these variables on the basis of the type of information they provide for building cases prediction model.
- We study the state-wise shift in spending pattern of citizens' in response to the pandemic by analyzing the customer response variables that are created using *Mastercard Transaction dataset*. We analyze the impact of Government's quarantine rules on these variables and note that the states that showed shift to online and contactless modes of payment were able to control the spread of the disease since these methods allow greater social distancing.
- We build a regression model to predict the new cases emerging in the following week at county-level and hence the number of extra medical resources that must be kept ready in advance to avoid critical situations. Thereafter we show the importance of these variables learned by the model in predicting the upcoming number of cases.
- We generate counterfactual examples by tuning the customer response variables to identify the county-specific measures that could help in reducing the number of expected cases in upcoming weeks. We also give estimate of the new overload of medical resources that would be

[1]Source: https://www.worldometers.info/coronavirus/

DOI 10.1109/ICMLA51294.2020.00209

required if these measures are enforced in order to flatten the curve of the pandemic.

II. Related work

Many review papers have been published for different applications of Machine Learning in studying COVID-19 like diagnosis, patient outcome prediction, tracking and predicting the spread, effect of quarantine, drug development, fake news prediction,etc.[5] We will focus on the applications of new cases prediction and effect of quarantine control that are similar to our task.

A. Quarantine control

Existing models study the impact of quarantine and travel restrictions on the spread of Covid-19 either using parameters based on historical data from SARS/MERS coronavirus epidemics [11] or are not implemented worldwide. [12] uses real-time mobility data and detailed case data including travel history from Wuhan, China to study the role of case importation on transmission in cities across China and analyze the impact of quarantine measures. [3] uses cellular mobility data from USA during 2019 and 2020 to demonstrate that there has been a substantial increase in social distancing since the start of the pandemic. Rates of social distancing varies by county characteristics. [2] use publicly available data using a mixed first-principles epidemiological equations and data-driven neural network model to indicate that the regions in which rapid government interventions and strict quarantine and isolation measures were implemented were successful in containing the spread of infection and prevent it from exploding exponentially, while [13] analyzed the impact of quarantine in various parts of the world. Some authors also described the impact of public policies on individuals' behavior. [36] studied individuals' changing behaviors in response to Government measures by designing a high performance environment for planning and responding in the event of epidemics; [37] focuses on designing a system that determines the duration of patient's stay at a hospital and identifies the medical resources required in order to avoid delays and increase efficiency in treatment while [38] studied the evolution of Lean Healthcare to increase the efficiency in treatments.

We use the *Mastercard Transaction dataset* to analyze the pattern of purchase followed by the citizens before and after the quarantine rules were enforced by the Government. We infer that the citizen response variables that we created using this financial data have a correlation with the number of COVID cases. We have not found any work that has focused on finding the impact of community financial activity on the spread of the disease.

B. Case prediction

In the prediction of confirmed cases using regression methods, two approaches have been used widely: (i)Many authors like [14], [15], [16], [17] focus on learning a logistic curve to predict the number of confirmed cases. (ii)Some works like [18], [19], [34], [11] predict the next day confirmed cases using m previous days confirmed cases. These works focus on using time-series algorithms like ARIMA to estimate COVID-19 spread[5].

Various types of deep learning neural networks having large number of hidden layers have been applied to predict the spread of Covid-19. [20], [21], [22], [23], [24] and [25] analyzed the use of LSTM, GRU, CNN and multilayer perceptons to estimate the number of confirmed cases in India.

The networks or graphs consisting of nodes and edges have also been used to study the spread of the infectious diseases as in [26], [27], [28] and [29]. When one of the nodes in a network gets infected, it can infect other nodes which are connected to the infected node. This process continues and the disease is spread to all the other nodes of the network.

Social media and search queries data-based methods like Qin et al. [31] predict the number of Covid-19 cases by using social media search indexes of Covid-19 symptoms. Jahanbin and Rahmanian [32] show that tweets extracted from twitter can be used in estimation of COVID-19 spread.

Along with the spread prediction using regression analysis, our work also focuses on finding the factors specific to a county that might be responsible for an increase in the number of new cases. We also estimate the reduction in the required medical resources if these factors are controlled.

III. Data collection

In this section, we describe the datasets that we have used for analyzing the citizen response to the pandemic control guidelines and spread prediction. We have used six different datasets to create the predictive variables for our spread prediction model. We merge a variety of data provided by these datasets to capture most of the information that is available at county level regarding the spread of the COVID, control measures taken by the officials, citizens' response to these measures and the impact of the response.

- *Mastercard transaction dataset* : Mastercard transaction data for USA[2] is aggregated at country, state and county level for 21 weeks starting from February 2020 (when COVID-19 cases started to rise in USA) to June 2020. Data consists of the following fields corresponding to each state/county: ratio of transactions with contact-based payment methods over contact-less methods, ratio of in-store purchase over online purchase, ratio of cross-state(county) purchase over domestic purchase and ratio of transactions at non-essential industries over essential industries.
- *Control measures dataset* : The Kaggle dataset[3] contains the information on the timeline of quarantine measures and "stay at home" instructions that were enforced by the Government.

[2]This dataset is a random sample of USA population since it incorporates information from only credit and debit card payments and does not include any information from cash payments and other payment modes.

[3]https://www.kaggle.com/lin0li/us-lockdown-dates-dataset

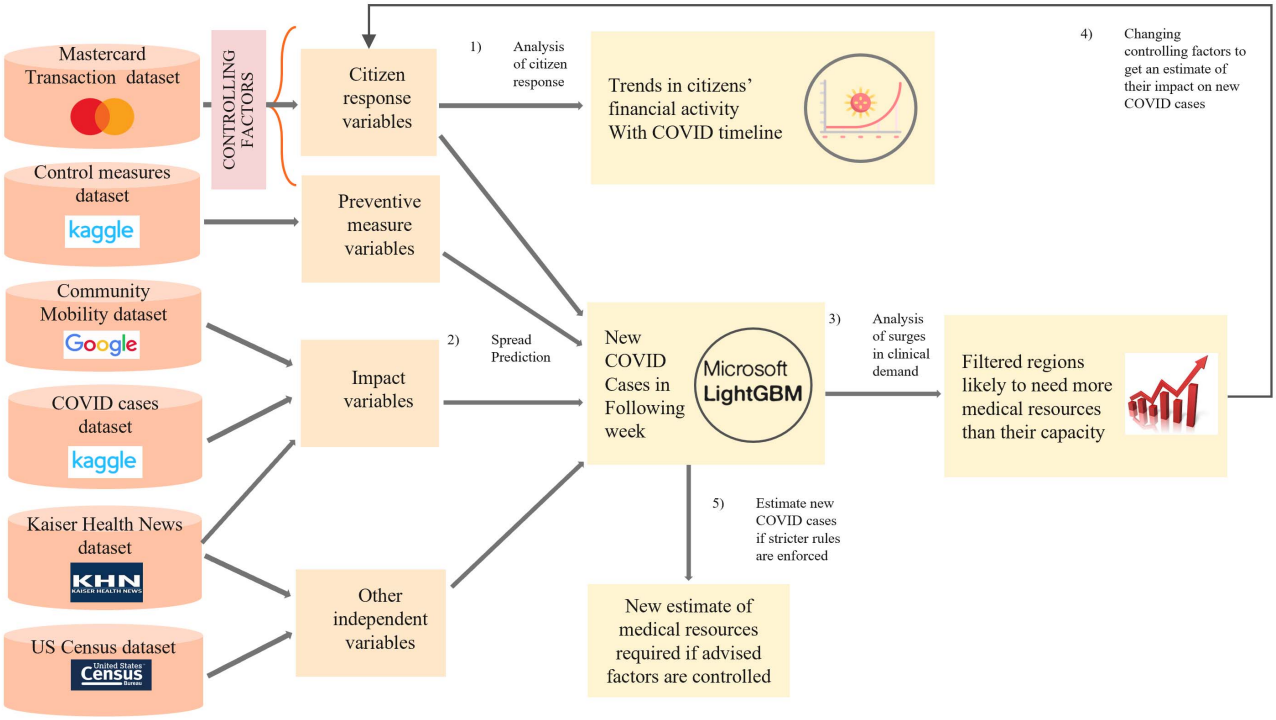

Figure. 1: Schematic representation of our proposed method. Six datasets shown are merged to create four types of variables. 1. Citizen response variables are analyzed 2. LightGBM model is trained to predict new COVID cases 3. Counties estimated to show rise in spread of COVID are analyzed 4. Citizen response variables are tuned to create counterfactual data examples. 5. New rise in cases is estimated on the counterfactual examples

- *Community Mobility dataset* : This contains Google Community Mobility Reports[4] on movement patterns over time by county, across different categories of places to design variables that indicate the percentage change on movement patterns from baseline.
- *US Census dataset* : US Census data [5] contains information about demographics like population, age, etc.
- *Kaiser Health News data* : The healthcare dataset[6] with information on availability of ICU beds in different regions.
- *COVID cases dataset* : The COVID dataset[7] contains information about daily confirmed cases, recovered cases and deaths due to COVID at county level.

IV. Methods

In this section, we describe the creation of multi-source dataset using the six datasets described in previous section. Further, we discuss the predictive modelling for predicting the COVID spread(number of new cases in the following week at county-level) and the generation of counterfactual examples to highlight the factors that need attention in order to flatten the curve of the spread. Figure. 1 shows the complete architecture of our method including the creation of variables from different datasets, analysis of citizens' response, predictive modelling and the generation of counterfactual examples by tuning the citizen response variables.

A. Variable Creation

We describe how different variables are created for the spread prediction model using six different datasets. We design four types of variables on the basis of the information they provide. Table 1 shows the statistical measures of these variables.

1) Citizen response variables: These are based on citizens' response to quarantine measures in terms of their spending pattern. We use the *Mastercard transaction data* for USA consisting of the following fields: citizen county, merchant county, industry and mode of transaction. We study the shift of citizens' spending pattern towards online and contactless technology after the quarantine announcements were made in their respective counties using 5 variables for each county from the time of emergence of first COVID positive case. These variables were created weekly by aggregating at county level: (i) ratio of in-store purchase to online purchase (OFF:ON); (ii) ratio of cross-county purchase to domestic purchase (XS:DOM);

[4]Mobility data can be found at: https://www.google.com/covid19/mobility/.
[5]https://www.census.gov/programs-surveys/popest.html
[6]https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/lookup
[7]Data can be accessed through: https://github.com/nytimes/covid-19-data/blob/master/.

TABLE I: Statistical measures of the variables

	Mean	SD	Min	25th Percentile	50th Percentile	75th Percentile	Max
Citizen response variables							
XS:DOM	0.615000	0.6460	0.0350	0.387000	0.549000	0.753000	40.800000
OFF:ON	2.390000	0.6740	0.5950	1.890000	2.350000	2.830000	31.500000
DIS:ESS	0.468000	0.1710	0.0506	0.347000	0.450000	0.567000	1.600000
CON:CONLS	270.000000	416.0000	5.8200	83.400000	162.000000	306.000000	16100.000000
XS DID:XS ESS	0.623000	0.2230	0.1090	0.461000	0.589000	0.748000	2.030000
Preventive measure variables							
CASES_LKDWN	5.430000	3.7500	0.0000	2.000000	6.000000	7.000000	12.000000
ICASE&LKDWN_LAG	3910.000000	12500.0000	-65.0000	-9.000000	-3.000000	5.000000	43900.000000
Impact variables							
%_RET&RECREATION	-12.500000	16.5000	-100.0000	-15.900000	-12.500000	-4.500000	150.000000
%_GRO&PHA	2.090000	10.4000	-67.0000	0.857000	2.090000	3.860000	147.000000
%_PARKS	22.600000	24.2000	-78.9000	22.600000	22.600000	22.600000	410.000000
%_WORKPLACES	-23.100000	13.6000	-77.2000	-31.200000	-23.100000	-18.300000	16.500000
%_TRANSIT_ST	-13.027146	13.3091	-91.0000	-13.027146	-13.027146	-13.027146	119.428571
%_RES	9.140000	4.3500	-4.0000	9.140000	9.140000	9.140000	32.000000
NUM_DEATHS	7.450000	55.6000	-8.0000	0.000000	0.000000	1.000000	2330.000000
%_BEDS_OCC	0.056000	0.0517	0.0000	0.027400	0.040800	0.066900	0.387000
Other independent variables							
STORE_POP_DENSITY	146.000000	108.0000	1.8600	136.000000	146.000000	146.000000	3060.000000
STORE_TXN_DENSITY	592.000000	2960.0000	0.0000	8.470000	46.900000	194.000000	62900.000000
NUM_BEDS	2.690000	0.6280	1.6000	2.300000	2.600000	3.100000	4.800000

(iii) ratio of non-essential or discretionary purchase to essential purchase (DIS:ESS); (iv) ratio of cross-county discretionary purchase to cross-county essential purchase (XS DIS:XS ESS); and (v) ratio of purchase using contact-based technology to contactless technology (CON:CONLS).

2) Preventive measure variables: These variables are based on the measures adopted by the Government officials to control the spread. Based on the quarantine measures taken by the county Government, we create two variables that contribute to the information on preventive measures taken locally using *Control measures dataset*. First, we create a variable for the number of positive COVID cases on the day when the shelter in place or stay at home announcements were made (CASES_LKDWN). Second, we create a variable for the lag between the first positive case and the quarantine announcement (ICASE&LKDWN_LAG).

3) Impact variables: These variables are based on the impact of preventive measures and citizen response on various industries after the quarantine announcements were made. According to public health officials, aggregated, anonymized insights of community mobility could be helpful to make critical decisions to combat COVID-19. These aim to provide insights into what has changed in response to policies aimed at combating COVID-19. So we use *Community Mobility dataset* to get information on movement patterns over time by county, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. For each of these categories we have a variable that indicates the percentage change from baseline (%_RET&RECREATION, %_GRO&PHA, %_PARKS, %_TRANSIT_ST, %_WORKPLACES,%_RES). We aggregate these variables weekly at county level. We also create a variable to represent the number of deaths (NUM_DEATHS) in the given week using *COVID cases dataset* and a variable to represent the beds already occupied by COVID patients in a county (%_BEDS_OCC) to capture impact on resource availability using *Kaiser Health News dataset*, totalling to 8 impact variables.

4) Other independent variables: These variables are fixed for a region and are based on the degree of social distancing that the region is able to permit in terms of the resources and population. We use *US Census dataset* to create two variables to capture population density of the county. These are population density wrt the number of stores present in the county (STORE_POP_DENSITY) and the number of transactions that are processed per store (STORE_TXN_DENSITY) in the respective counties. We also use *Kaiser Health News dataset* to create a variable for total ICU beds in the given county (NUM_BEDS) to capture resource availability.

B. Predictive Modelling

All the four types of variables mentioned in previous subsection contain predictive information to predict the spread of COVID. Figure. 1 shows how different data sources are combined to form these variables. We stack all of these 18 variables aggregated weekly at county level for 21 weeks (from February 2020 to June 2020). The target variable for modelling is the number of confirmed COVID cases in the next week that is obtained through COVID cases dataset. We perform regression analysis as described in Figure 1 to predict the level of participation in the new cases emerging in following week using LightGBM[39].

C. Generation of counterfactual examples

On the basis of the estimated COVID spread through modelling, we focus on finding the medical resources that might be needed in order to avoid delay in treatments and to ensure the availability of resources. Based on the number of ICU beds required and the ICU beds available in the county, we intend to give an estimate of the overload(if any) of the beds

TABLE II: State-wise analysis of impact of quarantine. The states highlighted in Red were not able to contain the spread of the COVID whereas the states highlighted in green were very responsive to the Government guidelines and hence were able to contain the spread. Other states highlighted in orange and lime showed average response.

	Impact of quarantine on spending pattern					Impact of uplifting the quarantine on spending pattern				
States	***XS:DOM***	***OFF:ON***	***DIS:ESS***	***CON:CONLS***	***XS DIS:XS ESS***	***XS:DOM***	***OFF:ON***	***DIS:ESS***	***CON:CONLS***	***XS DIS:XS ESS***
TX	***0.77***	***1.15***	***0.67***	***1.01***	***0.54***	***1.17***	***1.06***	***1.01***	***0.48***	***1.11***
FL	***0.75***	***1.19***	***0.69***	***0.92***	***0.46***	***1.21***	***1.04***	***0.60***	***0.39***	***1.13***
NY	***0.68***	***1.11***	***0.61***	***0.94***	***0.41***	***1.12***	***1.13***	***1.01***	***0.35***	***1.08***
CA	***0.68***	***1.16***	***0.55***	***0.97***	***0.46***	***1.13***	***0.86***	***0.94***	***0.37***	***1.08***
AZ	***0.73***	***1.24***	***0.67***	***0.93***	***0.50***	***1.16***	***0.91***	***0.93***	***0.40***	***1.07***
AK	***0.67***	***1.12***	***0.70***	***1.13***	***0.53***	***1.24***	***0.98***	***1.12***	***0.50***	***1.07***
DC	***0.95***	***1.29***	***0.54***	***0.85***	***0.26***	***1.14***	***0.73***	***1.01***	***0.29***	***1.14***
MN	***0.69***	***1.18***	***0.64***	***1.15***	***0.43***	***1.14***	***0.80***	***1.00***	***0.37***	***1.08***
RI	***0.68***	***1.10***	***0.65***	***0.92***	***0.55***	***1.15***	***0.88***	***0.99***	***0.49***	***1.12***

XS:DOM - ratio of cross-state purchase to domestic purchase
OFF:ON - ratio of in-store purchase to online purchase
DIS:ESS - ratio of non-essential purchase to essential purchase
CON:CONLS - ratio of purchase using contact-based technology to contactless technology
XS DIS:XS ESS - ratio of cross-state discretionary purchase to domestic discretionary purchase

that might be needed to avoid critical situations. We tune the citizen response variables since they are the only controlling variables decided by the response of citizens to the situation and can be controlled by enforcing stricter regulations on social distancing. We perform some experiments by modifying the values of these variables and analyze the changes in new predicted cases and hence the ICU beds required. This perturbation is in certain threshold range so that changes in new values of these variables are realistic. We generate Guassian noise in range *(0, feature_value/3)* and subtract this noise by original value. Thereafter we calculate already occupied beds by the product of accumulated active cases in the previous week and ICU admission rate which is 2.3%[8] for entire USA. Active cases till previous week are calculated by subtracting recovered cases and deaths from the total cases. We assume recovery rate for a county to be same as the recovery rate for the state in which a county lies.

V. Analysis and Results

In this section, we discuss the analysis we performed on the USA data at country, state and county level and the inferences that we made through our analysis. We also analyze the impact of Government regulations on citizen response at state level. Further, we discuss the analysis and results of the predictive modelling and some of the counterfactual examples that we generated.

A. Exploratory Data Analysis

As the step 1 of Figure. 1, we analysed the aggregated transaction data and found that after the second week of March 2020 (when the stay at home announcements were made), people shifted to online purchase of only essential items prioritizing their needs over wants. Discretionary purchase was reduced by 63% and in-store purchase was reduced by 42% as compared to the figures from March 2019. Figure 3 shows the trend in spending pattern of the citizens. We use the citizen response variables to understand the impact of regulatory initiatives of Government on the spending pattern of citizens. We observed positive slopes in our plotted curves during week 5 to week 8 which is the timeline when quarantine announcements were made. This positive slope indicates that citizens shifted to contactless technology and online payment methods over traditional payment methods that could have encouraged the spread of the disease. Also, we observed a rise in traditional payment methods again after week 12 due to quarantine upliftment in many regions. But the percentage of rise was low which shows that many people adapted to these practices that encourage more social distancing.

Almost every state had shown the decline in discretionary and in-store purchase but the percentage of decline varies from state to state; with District of Columbia (DC) showing the maximum decline of 53% and Texas(TX) showing the decline of only 22% in discretionary purchase. DC has also been the most responsive in terms of decline in in-store purchase. The COVID spread curve also suggests that the regions like DC and RI were able to contain the spread of the disease whereas TX and FL were very slow in response. Surprisingly DC shows only 2% decline in cross-state purchase. One reason behind this pattern might be its high dependency on other states for stores since DC accounts for only 0.2% of all the stores in USA, being the least contributor of stores. We analyzed the states based on their response to the Government advisory and the degree of containment. We picked DC and eight states for detailed analysis with few states like NY and CA showing slow response whereas others like DC and RI showing great response and recovery. Table 2 shows the detailed analysis of these states. For each of these states, we observed the ratio of citizen response variables before quarantine to after the quarantine for studying the impact of quarantine on the spending pattern of people. Similarly, we observed the ratio of citizen response variables after quarantine to after the quarantine was uplifted for studying the impact of quarantine upliftment on the spending pattern.

[8]Share of U.S. COVID-19 patients admitted to ICU, Jan.-May, 2020, by age can be found at https://www.statista.com/statistics/1127623/covid-19-patients-share-admitted-to-icu-us/

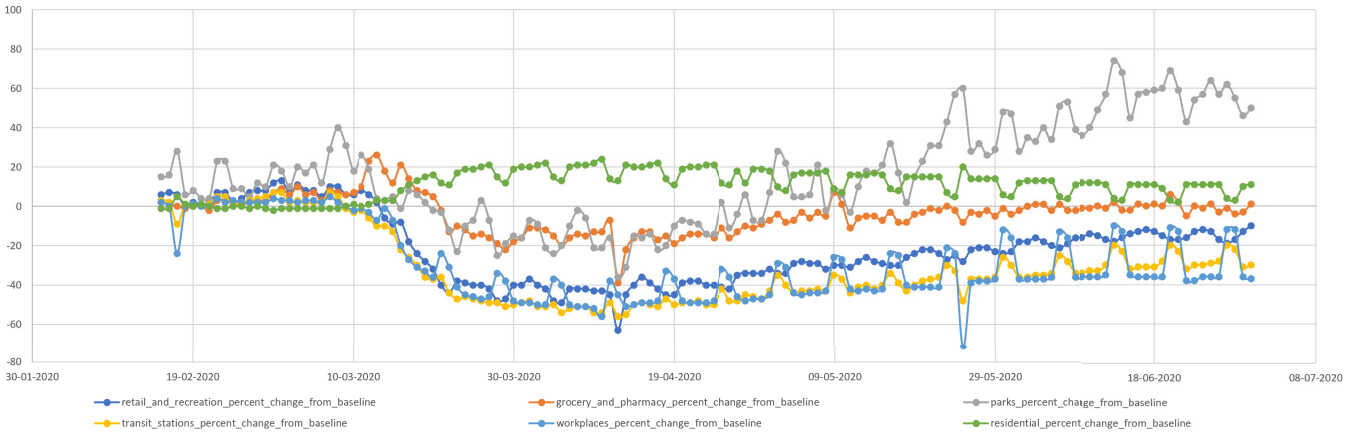

Figure. 2: This shows how visitors to categorized places changed after COVID cases increased in US as compared to baseline days. A baseline day is the normal value for that day of the week during January 2020 to February 2020

We observe that the states that were successful in disease containment had lower ratios of citizen response variables suggesting their positive response to the Government advisory and adapting to purchase of only essential items from domestic stores using online or contactless payment methods. We also observe that after uplifting the quarantine the scores for XS:DOM and DIS:ESS rose since the cross-state movements started and non-essential industries re-opened, but scores for OFF:ON and CON:CONLS went down further indicating that people permanently adopted these methods that allow greater social distancing.

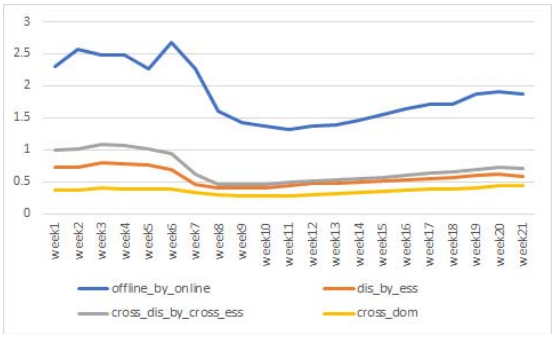

Figure. 3: Spending pattern of citizens with COVID timeline. Positive slopes during Week6 - Week8 (Enforcement of "stay at home") indicates a shift to online/contactless/essential/domestic purchase. Negative slopes during Week16-Week19 (Upliftment of "stay at home" orders) indicate that citizens resumed offline/contact-based/non-essential/cross-county purchase but with lower rates than earlier.

We also analyzed the impact variables derived from *Community Mobility Dataset* to study how the visitors to retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential changed with respect to the normal days before COVID spread in the US (Figure 2). For each of these categories we studied the percentage change from baseline. For workplaces (%_WORKPLACES), we observed that there is a decline of 23% from baseline. This can be directly correlated to work from home orders issued by companies. For transit stations (%_TRANSIT_ST) which includes places like public transport hubs such as subway, bus, and train stations, we saw a decline of 13% from baseline. For residential places (%_RES), we saw a minute change as compared to other categories as people already spend a lot of time at home (even on workdays). For parks and outdoor spaces (%_PARKS), we see spikes which represent large day-to-day changes. This is because visitors to parks are heavily influenced by the weather and holidays. We see an increase of 22% from baseline. This value drops during the period when the COVID impact was at its peak but after the mid week of May we can see an increasing trend. This can be related to the fact that people started doing activities to increase their immunity. For groceries and pharmacies (%_GRO&PHA), we see an increase of 2% from baseline. This value is not fluctuating much as this category encapsulates essential items. It drops at the peak week of COVID impact but increases after first week of May when people shifted towards online purchase of items. This increase can be justified by change in offline by online ratio. Both values OFF:ONN and %_GRO&PHA show a similar pattern. For retail and recreation (%_RET&RECREATION), we see 12% decline from baseline. This decline is due to strict policies enforced by the government to shut down all non essential services.

B. Significance of predictive variables

Figure. 3 shows the variable importance learnt by the predictive modelling. In the step 2 of Figure 1, we train the LightGBM model and found that the number of deaths in the given week is the most significant feature since it provides the information on the spread of COVID as well as the recovery rate. Other significant variables were the ratio of non-essential to essential purchase, ratio of cross-county

TABLE III: Resource overload reduction by generating counterfactual data. Highlighted(yellow) cells show the modified values of the variables that might reduce the rise in COVID spread. The counties where requirement of medical resources are estimated to rise beyond capacity are highlighted in red and the counties still in safe zone are highlighted in green.

County	ICU beds	Occupied beds	*XS:DOM*	*OFF:ON*	*DIS:ESS*	*CON:CONLS*	*XS DIS :XS ESS*	*Predicted cases*	*Beds required*	*Overload*	*New Overload*
Monroe	**1**	**1**	*0.60→0.41*	*2.83*	*0.54*	*176.95*	*0.79*	*2986→2908*	*69→67*	*-69*	*-67*
Fannin	**2**	**2**	*0.65→0.41*	*2.32*	*0.71*	*173.9*	*0.88*	*3479→3186*	*80→73*	*-80*	*-73*
Phelps	**18**	**0**	*0.49→0.10*	*2.46*	*0.48*	*240.2*	*0.62*	*980→946*	*23→22*	*-5*	*-4*
Blair	**50**	**1**	*0.265*	*2.55*	*0.32→0.24*	*76.25*	*0.46*	*1733→1631*	*40→38*	*+9*	*+11*
Jackson	**51**	**3**	*0.36*	*2.39*	*0.48→0.3*	*117*	*0.69*	*1449→1291*	*33→30*	*+15*	*+18*

to domestic purchase and store population density. The ratios contain the information about the response of citizens to the social distancing rules whereas the store population density represents the degree of social distancing that the region is able to permit, given that people follow in-store purchase.

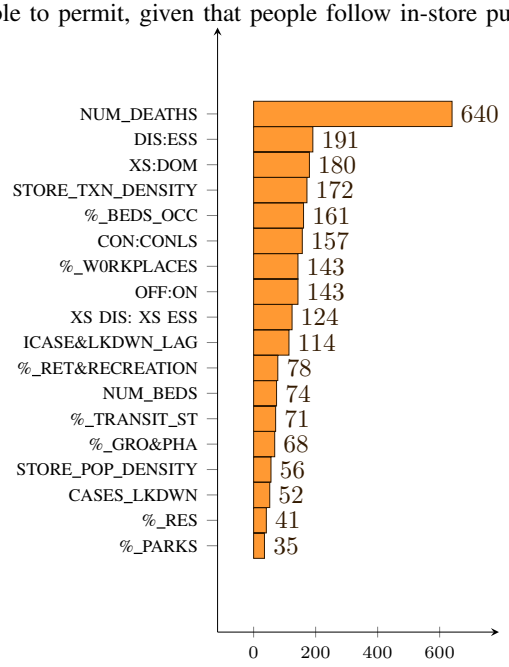

Figure. 4: Variable importance based on LightGBM; IN-TIME R^2 = 0.84; OUT-OF-TIME R^2 = 0.76

C. Counterfactual examples

Significance of the variables gives an overview of the factors that are affecting the spread of the disease. Table 3 shows the counterfactual data that we generated by modifying the values of the citizen response variables as part of step 3, 4 and 5 of Figure 1; and its impact on the expected reduction in number of new cases and hence the reduction in number of medical resources required in near future. We observed that if we reduce the values of these variables, we could reduce the number of new cases emerging in the following week and hence the number of ICU beds required. The reduction in spread due to reducing a variable might indicate an issue that needs to be given attention in order to flatten the curve. For an instance, the reduction in required number of ICU beds due to reduction in value of XS:DOM might indicate a need to re-establish domestic supply chain in order to control cross-county purchases.

VI. Discussion

Our analysis shows that the predictive variables we created contain significant amount of information to predict the spread of the disease in upcoming weeks. Using citizen response variables, we analyzed the impact of Government advisory on the pattern observed in citizens' purchase. Table 2 showed that the regions having lower value of these variables were able to flatten the spread curve after the quarantine rules were enforced. It also showed that the values for XS:DOM and DIS:ESS increased after upliftment of quarantine but the values for OFF:ON and CON:CONLS kept declining suggesting that people adopted online and contactless modes of payment that ensured greater social distancing.

We also designed a model to predict the number of COVID cases in the following week with R^2 value of 0.76. We found the importance of these variables learned by the model in estimating the new cases with the most important variables being the number of deaths, ratio of non-essential to essential purchase, ratio of cross-county to domestic purchase and store transaction density of the county. More information about recovery rate and percentage of medical resources required can improve the model performance and give more insights for healthcare preparedness.

We also try to produce counterfactual data examples by tuning customer response variables since they can be controlled by enforcing stricter policies. Table 3 gave an estimate of the new cases and the medical resources needed if value of these variables are reduced by enforcing stricter policies. This study gives an estimate of the region-wise factors that need to be controlled to reduce the spread of the disease. This might help the health officials to take efficient decisions by targeting a particular region according to its needs without disturbing other regions that are already following all possible measures of social distancing.

References

[1] Yan, Li & Zhang, Hai-Tao Xiao, Yang & Wang, Maolin & Sun, Chuan & Liang, Jing & Li, Shusheng & Zhang, Mingyang & Guo, Yuqi & Xiao, Ying & Cao, Haosen & Tan, Xi & Huang, Niannian & Jiao, Bo & Luo, Ailin & Cao, Zhiguo & Xu, Hui & Yuan, Ye. (2020). "Prediction of survival for severe Covid-19 patients with three clinical features:

development of a machine learning-based prognostic model with clinical data in Wuhan." 10.1101/2020.02.27.20028027.

[2] Dandekar, Raj & Barbastathis, George. (2020),"Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning." 10.1101/2020.04.03.20052084.

[3] Andersen, Martin. (2020), "Early Evidence on Social Distancing in Response to COVID-19 in the United States," *SSRN Electronic Journal.* 10.2139/ssrn.3569368.

[4] Ziff, Robert & Ziff, Anna. (2020). "Fractal kinetics of COVID-19 pandemic." 10.1101/2020.02.16.20023820.

[5] Amir Ahmad, Sunita Garhwal, Santosh Kumar Ray, "The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges," URL https://arxiv.org/pdf/2006.09184.pdf

[6] Vaishya, Raju & Javaid, Mohd & Khan, Ibrahim & Haleem, Abid. (2020). "Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews." 14.10.1016/j.dsx.2020.04.012.

[7] Alazab, Moutaz & Awajan, Albara & Mesleh, Abdelwadood & Abraham, Ajith & Jatana, Vansh & Alhyari, Salah. (2020). "COVID-19 Prediction and Detection Using Deep Learning." *12. 168-181.*

[8] Weissman, Gary & Crane-Droesch, Andrew & Chivers, Corey & Luong, ThaiBinh & Hanish, Asaf & Levy, Michael & Lubken, Jason & Becker, Michael & Draugelis, Michael & Anesi, George & Brennan, Patrick & Christie, Jason & III, C. & Mikkelsen, Mark & Halpern, Scott. (2020), "Locally Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic." *Annals of Internal Medicine. 173.* 10.7326/M20-1260.

[9] Jahanbin, Kia & Rahmanian, Vahid. (2020), "Using twitter and web news mining to predict COVID-19 outbreak." *Asian Pacific Journal of Tropical Medicine. 13.* 10.4103/1995-7645.279651.

[10] Punn, Narinder & Sonbhadra, Sanjay & Agarwal, Sonali. (2020), "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms." 10.1101/2020.04.08.20057679.

[11] D. Liu, L. Clemente, C. Poirier, X. Ding, M. Chinazzi, J. T. Davis, A. Vespignani, M. Santillana, "A machine learning methodology for realtime forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models (2020)." arXiv:2004. 04019.

[12] Kraemer, Moritz & Yang, Chia-Hung & Gutierrez, Bernardo & Wu, Chieh-Hsi & Klein, Brennan & Pigott, David & Plessis, Louis & Faria, Nuno & Li, Ruoran & Hanage, William & Brownstein, John & Layan, Maylis & Vespignani, Alessandro & Tian, Huaiyu & Dye, Christopher & Pybus, Oliver & Scarpino, Samuel. (2020), "The effect of human mobility and control measures on the COVID-19 epidemic in China." *Science. 368. eabb4218.* 10.1126/science.abb4218.

[13] Singer, H. M, "Short-term predictions of country-specific Covid-19 infection rates based on power law scaling exponents." arXiv:2003.11997 (2020).

[14] Pandey, Gaurav. (2020), "SEIR and Regression Model-based COVID-19 outbreak predictions in India" (Preprint). 10.2196/preprints.19406.

[15] Gu, Chaolin & Zhu, Jie & Sun, Yifei & Zhou, Kai & Gu, Jiang. (2020), "The inflection point about COVID-19 may have passed." *Science Bulletin. 65.* 10.1016/j.scib.2020.02.025.

[16] Xu, Stanley & Clarke, Christina & Shetterly, Susan & Narwaney, Komal. (2020), "Estimating the Growth Rate and Doubling Time for Short-Term Prediction and Monitoring Trend During the COVID-19 Pandemic with a SAS Macro." 10.1101/2020.04.08.20057943.

[17] Li, Yi & Liang, Meng & Yin, Xianhong & Liu, Xiaoyu & Hao, Meng & Hu, Zixin & Wang, Yi & Jin, Li. (2020), "COVID-19 Epidemic Outside China: 34 Founders and Exponential Growth." 10.1101/2020.03.01.20029819.

[18] Gupta, Rajan & Pal, S.K.. (2020), "Trend Analysis and Forecasting of COVID-19 outbreak in India." 10.35543/osf.io/e547c.

[19] Chakraborty, Tanujit & Ghosh, Indrajit. (2020), "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis." 10.1101/2020.04.09.20059311.

[20] Tomar, Anuradha & Gupta, Neeraj. (2020), "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures." *Science of The Total Environment. 728. 138762.* 10.1016/j.scitotenv.2020.138762.

[21] Z. Hu, Q. Ge, S. Li, L. Jin, M. Xiong, "Artificial intelligence forecasting of covid-19 in china (2020)." arXiv:2002.07112.

[22] Charte, David & Charte, F. & García, Salvador & Del Jesus, María José & Herrera, Francisco. (2018), "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines" *Information Fusion. 44. 78-96.* 10.1016/j.inffus.2017.12.007.

[23] Sr, Samir & Dutta, Shawni. (2020), "Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release (Preprint)." 10.2196/preprints.19526.

[24] N. M. Ghazaly, M. A. Abdel-Fattah, A. A. A. El-Aziz, "Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network", *International Journal of Advanced Science and Technology 29 (5s).*

[25] Huang, Chiou-Jye & Chen, Yung-Hsiang & Ma, Yuxuan & Kuo, Ping-Huan. (2020), "Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China." 10.1101/2020.03.23.20041608.

[26] Zhuang, Zian & Zhao, Shi & Lin, Qianying & Cao, Pei-Hua & Lou, Yijun & Yang, Lin & Yang, Shu & He, Daihai & Xiao, Li. (2020). "Preliminary estimating the reproduction number of the coronavirus disease (COVID-19) outbreak in Republic of Korea and Italy by 5 March 2020." *International Journal of Infectious Diseases. 95.* 10.1016/j.ijid.2020.04.044.

[27] Herrmann, Helena & Schwartz, Jean-Marc. (2020), "Using network science to propose strategies for effectively dealing with pandemics: The COVID-19 example." 10.1101/2020.04.02.20050468.

[28] Pujari, Bhalchandra & Shekatkar, Snehal. (2020), "Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India." 10.1101/2020.03.13.20035386.

[29] K. S. Pokkuluri, N. D. Nedunuri, U. S. Usha, "A novel cellular automata classifier for covid-19 prediction", *Journal of Health Sciences 10 (1) (2020) 34–38.*

[30] C. Fan, T. Cai, Z. Gai, Y. Wu, "The relationship between the migrant population migration network and the risk of covid 19 transmission in china empirical analysis and prediction in prefecture level cities", *International Journal of Environmental Research and Public Health 17.*

[31] Qin, Lei & Sun, Qiang & Wang, Yidan & Wu, Ke-Fei & Chen, Mingchih & Shia, Ben-Chang & Wu, Szu-Yuan. (2020), "Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index." *International Journal of Environmental Research and Public Health. 17. 2365.* 10.3390/ijerph17072365.

[32] Jahanbin, Kia & Rahmanian, Vahid. (2020), "Using twitter and web news mining to predict COVID-19 outbreak." *Asian Pacific Journal of Tropical Medicine. 13.* 10.4103/1995-7645.279651.

[33] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, S. R Niakan Kalhori, "Predicting covid-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study", *JMIR Public Health Surveillance 6 (2) (2020) e18828.* doi:10.2196/18828.

[34] Benvenuto, Domenico & Giovanetti, Marta & Vassallo, Lazzaro & Angeletti, Silvia & Ciccozzi, Massimo. (2020), "Application of the ARIMA model on the COVID-2019 epidemic dataset.", *Data in Brief. 29. 105340.* 10.1016/j.dib.2020.105340.

[35] Deb, Soudeep, and Manidipa Majumdar, "A time series method to analyze incidence pattern and estimate reproduction number of COVID-19." arXiv preprint arXiv:2003.10655 (2020).

[36] Bisset, Keith & Eubank, Stephen & Marathe, Madhav. (2012), "High performance informatics for pandemic preparedness." *Proceedings - Winter Simulation Conference. 1-12.* 10.1109/WSC.2012.6465211.

[37] Patvivatsiri, Lisa & Jr, Elliot & Xi, Ouyang. (2007), "Modeling bioterrorism preparedness with simulation in rural healthcare system." *1155-1160.* 10.1109/WSC.2007.4419716.

[38] Zepeda-Lugo, Carlos & Tlapa, Diego & Báez, Yolanda & Limon, Jorge. (2018), "Critical Factors of Lean Healthcare: an Overview." *1-7.* 10.1145/3242789.3242837.

[39] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., "LightGBM: a highly efficient gradient boosting decision tree" *in Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp. 3149-3157, 2017.*

[40] Petrosillo, Nicola & Viceconte, Giulio & Ergonul, Onder & Ippolito, Giuseppe & Petersen, Eskild. (2020), "COVID-19, SARS and MERS: are they closely related?." *Clinical Microbiology and Infection. 26.* 10.1016/j.cmi.2020.03.026.