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) which began during December 2019. More than 113 million cases have been reported across 219 countries and territories as of 27 February 2021, resulting in more than 2.5 million deaths.[1] Authorities worldwide have responded by implementing travel restrictions, home quarantine, workplace hazard controls, and facility closures. 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 [2], we aim to analyze the pattern in citizens' movement post the quarantine orders so that the officials can make efficient decisions. 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[2] or are not implemented worldwide[3][4]. Pandemics are rare and have different characteristics so the data of other coronaviruses cannot be used for Covid-19[5]. 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. Many authors like [6],[7],[8],[9] focus on learning a logistic curve to predict the number of confirmed cases. We also present a machine learning model to predict new cases in following week to notify the health workers about the medical resources that might be needed in future. The further analysis gives insights about the impact of control policies 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 study the state-wise shift in spending pattern of citizens' in response to the Government's quarantine rules by analyzing the customer response variables that are created using *Mastercard Transaction dataset*.
- 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.
- 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 required if these measures are enforced in order to flatten the curve of the pandemic.

II. Data collection

We use six different data sources to create the predictive variables for our spread prediction model.

- *Mastercard transaction dataset* : Mastercard transaction data for USA[3] 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

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

[2]This paper was published in 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)[1]. The full paper can be found at https://ieeexplore.ieee.org/document/9356320

[3]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.

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[4] contains the information on the timeline of quarantine measures and "stay at home" instructions that were enforced by the Government.
- *Community Mobility dataset* : This contains Google Community Mobility Reports[5] 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 [6] contains information about demographics like population, age, etc.
- *Kaiser Health News data* : The healthcare dataset[7] with information on availability of ICU beds in different regions.
- *COVID cases dataset* : The COVID dataset[8] contains information about daily confirmed cases, recovered cases and deaths due to COVID at county level.

III. 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.

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.

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. The 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); (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. 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. 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 to predict the level of participation in the new cases emerging in following week using LightGBM[10].

C. Generation of counterfactual examples

On the basis of the estimated COVID spread, we focus on finding the medical resources that might be needed in order

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

[5]Mobility data can be found at: https://www.google.com/covid19/mobility/.

[6]https://www.census.gov/programs-surveys/popest.html

[7]https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/lookup

[8]Data can be accessed through: https://github.com/nytimes/covid-19-data/blob/master/.

TABLE I: 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

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 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 Gaussian 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%[9] 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.

IV. 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

First, 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 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. 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 1 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.

[9]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/

TABLE II: 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*

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.

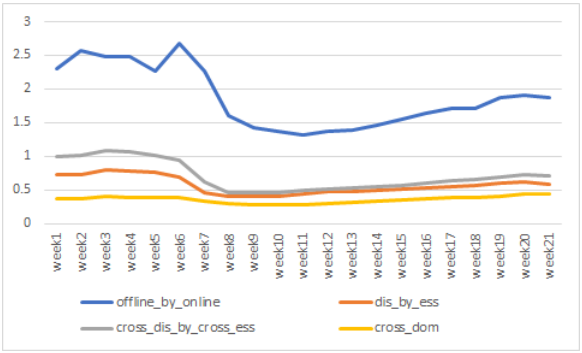

Figure. 1: Spending pattern of citizens with COVID timeline. Positive slopes during Week6 - Week8 (Enforcement of "stay at home") indicates a shift to online/contactless/essential/domestic purchase. Negative slopes during Week16-Week19 (Upliftment of "stay at home" orders) indicate that citizens resumed 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 different places changed with respect to the normal days before COVID spread in the US. For each of these categories we studied the percentage change from baseline. For workplaces (%_WORKPLACES), we observed a decline of 23% from baseline that can be directly correlated to work from home orders issued by companies. For transit stations (%_TRANSIT_ST), 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. 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. 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

The predictive model achieved in-time R^2 of 0.84 and out-of-time R^2 of 0.76. The most significant variables learnt by this model were number of deaths, ratio of non-essential to essential purchase, ratio of cross-county 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.

C. Counterfactual examples

Table 2 shows the counterfactual data that we generated by modifying the values of the citizen response variables; 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. 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.

V. Conclusion

In this paper, we analyzed the impact of Government advisory on the citizens' movement. We found various factors that can affect the COVID-19 spread and designed 18 variables including citizen response variables, impact variables, mobility variables and preventive measures' variables using various data sources. We designed a machine learning model to predict the number of COVID cases in the following week using these variables. The model found these variables to be highly significant. We analysed the region-wise factors that need to be controlled to reduce the spread of the disease. This would permit the officials to take efficient decisions by targeting only the regions that require attention.

REFERENCES

[1] N. Mulay, V. Bishnoi, H. Charotia, S. Asthana, G. Dhama and A. Arora, "Pandemic spread prediction and healthcare preparedness through financial and mobility data," *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2020, pp. 1340-1347*, doi: 10.1109/ICMLA51294.2020.00209.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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