How predictive models can aid in the battle against COVID-19

3 April 2020

As the COVID-19 pandemic grows, KPMG in India is developing machine-learning models to predict the severity of the outbreak and identify at-risk populations across the country by state and eventually, by district.

By Sanjaya Sharma, Senior Advisor, Risk Consulting-IT Advisory, KPMG in India

Artificial Intelligence-based predictive modelling is emerging as a powerful weapon in humanity's fight against disease outbreaks. They not only help us identify vulnerable and infected individuals, but also help us to predict and understand the severity and geographical spread of the contagion.

In 'The Rules of Contagion', infectious disease expert Adam Kucharski writes that the 'reproduction number' of a disease, or the number of new infections a typical infectious person generates on average, depends on four factors, or what he calls DOTS for short: 1) the duration of time a person is infectious 2) the average number of opportunities they have to spread the infection each day 3) transmission or the probability of an opportunity translating into actual infection 4) susceptibility of the population which can be gauged by demographic profiles, medical histories, etc.

The government's stringent measures, including the 21-day lockdown, can help limit the opportunities and reduce the probability of transmission. However, the other factors at play, i.e., the duration and susceptibility, cannot be controlled.

At the time of writing, India has about 1466 confirmed cases of COVID-19, spread across 27 states and union territories, according to official government data. As the COVID-19 pandemic grows, KPMG in India is developing machine-learning models to predict the severity of the outbreak and identify at-risk populations across the country by state and eventually, by district. To achieve this, we are building the following models:

- **Individual testing:** This binary classification model aims to predict whether a person will be tested positive or negative based on data from samples taken. Using data from the official online COVID-19 database, we looked at 3552 cases of which about 580 tested positive. Results based on the training data revealed that our model could predict positive cases with a precision of 72 per cent (the total number of cases) and recall of about 68 per cent (the actual individuals who caught the infection). For negative cases, the accuracy of the results crossed 90 per cent on both measures (precision and recall).

- **Positive cases per state:** This model can predict the number of positive cases that are expected everyday using state-wise, timeseries data. The following table compares our model's predictions for new positive cases on a daily basis across five states, with actual positive cases. In Punjab, for instance, the actual cases fell from four on March 25 to zero on March 26; our model, too, has predicted this. The model has largely captured the directional changes in the number of new positive cases.

	Prediction Window	Predicted Positive Cases	Actual Positive Cases	MAE
Punjab	26 to 28 March	10.3	9	1.29
West Bengal	26 to 28 March	6.3	6	0.32
Karnataka	26 to 28 March	26.0	28	2.00
UP	26 to 28 March	8.2	16	7.82

As can be seen in the chart above, the model has been able to make fairly accurate predictions for three of the four states indicated, basis the available data.

While the accuracy of our initial predictions offers encouragement, we are expecting significant improvement as the dataset expands. We are hopeful that by the second week of April, we can develop an integrated model that will—to a reasonable degree of accuracy—predict and quantify the severity of contagion in each geographical state, on an ongoing basis. This model can later be extended to analyse district-level trends.

We are developing processes to incorporate live data into our model, reducing the need for manual intervention. Live dashboards, meanwhile, will help track potential geographical hotspots and estimates of the number of possible cases. By tagging this data to the availability of healthcare facilities and personal protective equipment in the area, we can identify regional gaps in requirements.

Several parts of India do not have adequate access to COVID-19 testing kits, as the world faces huge shortages amid surging demand. One useful application of our model could be in such regions. The model can help identify vulnerable individuals by region. Targeted measures to control the spread of the virus, for instance quarantining at-risk people, can potentially be introduced based on our model's projections. Our models can also be applied to specific communities, such as an apartment complex or a migrant camp, to help focus stage-wide testing.

The success of predictive models depends largely on data, time and even on smart coordination between stakeholders. Patterns can change quickly in a pandemic, as the COVID-19 outbreak has shown. A few days ago, for instance, some individuals were identifiable as being possible sources of infection since they had travelled to affected countries. However, with the ban on all travel, we need to bolster the datasets with contact tracing data and test data from migrant groups. For more accurate predictions, constant monitoring of the model's outputs and updated data are required.

Pandemic mathematics can also have wider practical applications, such as identifying financial disasters and predicting the fallout. As the world braves the worst of the COVID-19 pandemic, AI and ML modelling can help us be prepared for, and possibly effectively contain, future disease outbreaks.