Evaluating Factors Influencing COVID-19 Outcomes across Countries Using Decision Trees (Student Abstract)

Aniruddha Pokhrel*, Nikesh Subedi*, Saurav Keshari Aryal
Howard University
2400 6th St NW
Washington, District of Columbia 20059 USA
{aniruddha.pokhrel, nikesh.subedi} @bison.howard.edu, saurav.aryal@howard.edu

Abstract

While humanity prepares for a post-pandemic world and a return to normality through worldwide vaccination campaigns, each country experienced different levels of impact based on natural, political, regulatory, and socio-economic factors. To prepare for a possible future with COVID-19 and similar outbreaks, it is imperative to understand how each of these factors impacted spread and mortality. We train and tune two decision tree regression models to predict COVID-related cases and deaths using a multitude of features. Our findings suggest that, at the country-level, GDP per capita and comorbidity mortality rate are best predictors for both outcomes. Furthermore, latitude and smoking prevalence are also significantly related to COVID-related spread and mortality.

Introduction

Globally, as of 15 September 2022, there have been over 607 million cases resulting in 6.49 million deaths from COVID-19 reported to World Health Organization. The COVID-19 pandemic, caused by the SARS-CoV-2 virus, is highly transmissible through small liquid particles excreted through the mouth or nose. Unsurprisingly, the pandemic had a devastating effect on the global economy and daily life. While humanity prepares for a post-pandemic world and a return to normality through worldwide vaccination campaigns, each country experienced different levels of impact based on natural, political, regulatory, and socio-economic factors. To prepare for a possible future with COVID-19 and similar outbreaks, it is imperative to understand how each of these factors impacted spread and mortality.

Research has been done between the relationship between COVID-related outcomes (cases and deaths) and factors. In the following section, we explore relevant works in this domain and discuss their limitations. In succeeding sections, we discuss the dataset and approaches to modeling and evaluating feature importance. Finally, we outline the results of our analysis and discuss limitations and future work.

Relevant Works

Several factors and their relationship to COVID-19 were noted. The factors studied include: average temperature in celsius (Xie and Zhu 2020), comorbidity mortality rate (Xie et al. 2020), vaccinations (Chen et al. 2022), testing (Evans et al. 2021), diabetes prevalence (Peric and Stulnig 2020), elevation (Stephens, Chernyavskiy, and Bruns 2021), gross domestic product (GDP), GDP per capita usd, human capital index and latitude (Sajadi et al. 2020), population density (Bhadra, Mukherjee, and Sarkar 2021), and smoking prevalence (Patanavanich and Glantz 2020). In the interest of brevity, the specific approaches taken by each researcher is not detailed, but we urge interested readers to peruse the works cited above.

Upon a review of the relevant works, we noticed that most of the works studied the relationship between one or a few variables and COVID-related outcomes. Furthermore, the analysis was limited to a specific country. While this approach may serve to understand a country-specific factors and associated outcomes, the results may not be generalizable to other countries. Moreover, since only one or a few features were studied, the impact of each feature on COVID-related outcomes is not immediately clear.

While we note that the list of variables presented is not exhaustive and research is still ongoing, we hypothesize that using country-level differences in these factors and their associated outcomes a model to predict cumulative cases and deaths can help us understand the importance of each factor. We hope that this understanding that be used in the future to better prepare for any potential outbreaks.

Methodology

The data used for this project was obtained from Google Health's COVID-19 Open Data Repository (Wahltinez et al. 2020). The dataset includes publicly sourced time-series data which covers a multitude of factors for over 120 countries. Of these factors, we only utilize factors which have been covered in the relevant works section.

From the larger dataset, we first created two output variables: *cases* and *deaths* by scaling each countries cumulative cases/death by their population. Since the dataset represented time series and we only used numeric variables, all features that are cumulative were scaled by population and those which were non-cumulative were averaged. Since all data-points are positive-valued, missing values were then filled in with a discriminating, arbitary value of -99999.

We train two decision tree regression models (deaths and

*These authors contributed equally.

cases) (Pedregosa et al. 2011) since they enable interpretation and explanation of feature importance. The best-model was utilized to report our findings after tuning using 10-fold cross-validation with GridSearch on the entire dataset. Since other standard error metrics are harder to interpret for this problem, we opt to report Mean Absolute Percentage Error (MAPE). Feature importance is reported using two methods: node-level importance (the decrease in node impurity weighted by the probability of reaching that node) and feature-level permutation importance (the decrease in a model score when a single feature value is randomly shuffled). To ensure reproducibility, the preprocessed dataset, source code, tuned models, and hyperparameters have been provided as supplemental materials.

Results

Upon training and tuning both models, we achieved a MAPE of 2.48% and 2.95% for the cases and death proportion predictions. Next, we report feature importances using node-level importance (NI) and permutation importance (PI) for all features and both models in Table 1. Of the features studied, comorbidity mortality, GDP per capita, and latitude were best predictors for cases whereas comorbidity mortality, GDP per capita, and smoking prevalence were best predictors for deaths. It stands to be reasoned that the existence of pre-existing conditions (comorbidity mortality rate) and available resources per person (GDP per capita) are the best predictors for both cases and deaths. Moreover, we do find it interesting that latitude has a significant effect in spread and smoking prevalence is more directly related to covid-related deaths than more directly observed factors such as testing or vaccination rates. However, these results do not necessarily imply that the other factors do not significantly impact covid-related outcomes rather at the global scale comorbidity mortality rate and GDP per capita are the best predictors of survivability.

There are certain limitations to our findings. Although we utilized a well-established datasource, the source is secondary and contains missing data for a significant number of countries. Furthermore, research has shown countries such as: China, India, Brazil, France, Italy, United States, Turkey, Iran, and Spain have a large number of unreported or undetected cases. Since we summarized the data into means, we could not study the progression of COVID-19 over time. We were also not able to factor in other pertinent variables such as mutations and mobility which directly impact outcomes.

Conclusion

Using openly available data, we were able to train and tune a decision tree regression model to figure out the feature importance of a multitude of factors. We found that, at the country-level, GDP per capita and comorbidity mortality rates are best predictors for both deaths and cases. Additionally, latitude and smoking prevalance are also significant predictors for cases and deaths respectively. However, modeling progression of the disease over time and incorporating regulatory or natural factors should be future work.

Feature name	**Cases**		**Deaths**	
	NI	PI	NI	PI
temperature	0.000	0.000	0.000	0.000
comorbidity mortality	**0.582**	**0.543**	**0.110**	**0.135**
fully vaccinated	0.000	0.000	0.000	0.000
new vaccinated	0.000	0.000	0.060	0.060
new tested	0.055	0.091	0.000	0.091
diabetes	0.000	0.000	0.000	0.000
gdp per capita	**0.195**	**0.412**	**0.559**	**0.672**
gdp usd	0.000	0.000	0.000	0.000
human capital index	0.057	0.080	0.003	0.003
latitude	0.090	**0.112**	0.061	0.088
population density	0.000	0.000	0.000	0.000
smoking prevalence	0.021	0.028	**0.206**	**0.152**

Table 1: Importance of features for each model

References

Bhadra, A.; Mukherjee, A.; and Sarkar, K. 2021. Impact of population density on Covid-19 infected and mortality rate in India. *Modeling earth systems and environment*, 7(1): 623–629.

Chen, X.; Huang, H.; Ju, J.; Sun, R.; and Zhang, J. 2022. Impact of vaccination on the COVID-19 pandemic in US states. *Scientific reports*, 12(1): 1–10.

Evans, S.; Agnew, E.; Vynnycky, E.; Stimson, J.; Bhattacharya, A.; Rooney, C.; Warne, B.; and Robotham, J. 2021. The impact of testing and infection prevention and control strategies on within-hospital transmission dynamics of COVID-19 in English hospitals. *Philosophical Transactions of the Royal Society B*, 376(1829): 20200268.

Patanavanich, R.; and Glantz, S. A. 2020. Smoking is associated with COVID-19 progression: a meta-analysis. *Nicotine and tobacco research*, 22(9): 1653–1656.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Peric, S.; and Stulnig, T. M. 2020. Diabetes and COVID-19. *Wiener Klinische Wochenschrift*, 132(13): 356–361.

Sajadi, M. M.; Habibzadeh, P.; Vintzileos, A.; Shokouhi, S.; Miralles-Wilhelm, F.; and Amoroso, A. 2020. Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19. *Social Science Research Network*.

Stephens, K. E.; Chernyavskiy, P.; and Bruns, D. R. 2021. Impact of altitude on COVID-19 infection and death in the United States: A modeling and observational study. *PLoS One*, 16(1): e0245055.

Wahltinez, O.; et al. 2020. COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2. Work in progress.

Xie, J.; and Zhu, Y. 2020. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment*, 724: 138201.

Xie, Y.; You, Q.; Wu, C.; Cao, S.; Qu, G.; Yan, X.; Han, X.; Wang, C.; and Zhang, H. 2020. Impact of cardiovascular disease on clinical characteristics and outcomes of coronavirus disease 2019 (COVID-19). *Circulation Journal*, CJ–20.