Rosey’s comments for:

Lexi

Excellent job Lexi! I think this topic is really interesting and there are a lot of dynamics in your dataset. I like your EDA exploring the difference of trending categories in UK and US, and many insights can be drawn from there (Are people in UK care more about pets than people in US? Possible!). In terms of suggestions, I think the definition of “popularity” is the key, and specifically, the relationship between time and “popularity”.

If the discussion is about one point of time, the pros will be a lot of variables are comparable such as the number of views depending on the tags or categories; however, the cons will be it ignores the way that Youtube works because people’s attention can be shifted by a lot of external factors (such as COVID).

If the discussion is about a trend, it will reflect the actual situation better. But the challenge will be how to put each individual video on the same scale to measure its popularity. I noticed that you have tried some methods to define it (such as “view\_increase” and “like\_increase”, etc.), and I would recommend reflecting the “average increase rate” by calculating *(max-min)/ number of days* *trending* to mitigate the effect of time. Another issue to consider will be how to map COVID daily cases on this analysis. Maybe it will be helpful to include a dummy to signal if the increase rate of cases during the trending days exceeds a certain number.

Maryam

Great presentation Maryam! I like the way you approach this macro-level policy issue on education by quantifying both dependent and independent variables, and the V-Dem data set looks to fit your analysis perfectly. For the data preprocessing part, I think it’s a good practice to include lagged components in your dependent variable to fight autocorrelation. Good job!

In terms of suggestions, I think the major challenge is the poor performance of the models.

There could be several reasons. First, I guess it will be due to missingness. I noticed that you’re imputing the missing values by mean. Maybe you can use KNN imputation to reflect the dataset better. (But as in your original data set, there are not so many missing values as a proportion. So, I guess it’s not the major issue. ) Second, maybe even more important than missingness, is the skewed continuous variables. If countries are clustered on the high-end of the V-Dem index but have different results in the education outcome, there could be an issue. In response to that, I suggest recoding the continuous variables to put them into different bins depending on a reasonable scale. (i.e. more bins on the high-end of the index than low-end).

Yousuf

Good presentation Yousuf! The details of your projects are very clearly explained, and I think democracy backsliding is really an important topic in terms of socioeconomical development of a country given the whole world has experienced a super hard 2020 and needs to move forward. Your presentation actually gave me some important insights on my own project as my data also suffer the issue of rare events. (i.e. 95% 0 and 5% 1 for the dependent variable). The balanced accuracy score and confusion matrix you used are super helpful tools to analyze this issue. Thank you : ) I terms of suggestions, I think there is one major aspect you need to keep in mind. For the three different models you are considering using, they might not be directly comparable as the units of analysis are different. For the V-Dem data, the variables are index ranging from 0-1, but for the econ data from Gapminder and World Bank, some variables are index like, but others are absolute values. Standardizing the scales might be a way around it and I guess they should be given the same effect of time (i.e. take the same magnitude of lag). Another thing I’m concerned about is the full model. Will the underlying calculation for the index in the V-Dem dataset be highly correlated with the econ data you are trying to use? In that case, you might consider the effect of multicollinearity. Also, I believe interactions among the variables will provide a lot of insights. Looking forward to how the final results will look like!

Charlie

Good job Charlie! I think this is a very relevant topic given there is a lot of uncertainties around the world during the COVID time. I like how you build up your model into two main categories including state capacity and proper response makes a lot of sense. Using the rate of change instead of the absolute values is also a reasonable way to fight autocorrelation. In terms of suggestions, I guess the major issue would be the performance of the models in the machine learning part. For many variables such as the “hospital bed”, could be very stable over time, but cases are changing much more quickly. Possible solution could be scaling the hospital bed (take the log), discretize the value by setting a limit depending on the distribution, or change the weeks to dummies to investigate the fixed effect. Anyway, to capture the change on the same scale should be the key. My additional thoughts are maybe you can include more independent variables such as the institution of the country into your analysis and create some interaction variables : )

Matt

Nice presentation Matt. This is a really interesting and important topic! I think I actually learned a lot from your presentation in terms of data preprocessing such as constructing three different models including full, restricted and imputed data. I haven’t thought about this before, thank you for providing a way to think about how to capture the effect of preprocessing the data. In terms of suggestion, I wish to know more about the models. You used the fixed effects OLS(countries to dummies) in your analysis, will that be computational expensive? How will the results look like if you keep them as ids? Also, I’m curious about different types of imputation (mean, frequency, knn, etc.) I guess they will lead to different results. Besides, there could be other variables related to this issue that could necessarily be interesting such as considering the gender-based indicators including gender wage gap or marriage rate(which will influence the opportunity costs of working)