Determining the important features for estimating the reproduction number in the COVID-19 pandemic

Anshul Toshniwal, Kseniia Shevchenko, Savyaraj Deshmukh

***Abstract*—In less than a year the COVID-19 pandemic has disrupted the global social fabric and has caused a significant loss in life and resources. To mitigate the effects of COVID-19, countries need to evaluate the effect of various policies on the Reproduction number (R0). For this purpose, a deep learning tool called 'What If... simulator' based on a hybrid LSTM model has been developed. In this project, we will try to improve our understanding of this model by focusing on the impact of various temporal features on the model performance.**

I. Introduction

The efficacy of COVID-19 policies across countries is not universal and is dependent on many heterogeneous features of each country. The impact of policies needs to be evaluated by weighing the trade-off between saving lives and reducing the economic and social collateral damage. To guide our decision, it is imperative to have a probabilistic estimate of the impact of policies on the reproduction number to construct an optimal policy mix.

The model used in our project is described in the Master Thesis "Adaptive Mitigation: Identification of the Dynamic Drivers of Effective Policy during the COVID-19 Pandemic" by Thierry Bossy. [1] The model is a hybrid Neural Network which combined an Long-Short-Term Memory (LSTM) layer with a multilayer perceptron (MCP) and is used for prediction of the reproduction number. For generalizability a leave-one-out cross-validation is used. Prediction for a given country is performed after training on all other countries.

Certain steps have also been taken in terms of the feature selection. Thus, epidemiological data are not used because of possible distortion of the prediction results due to incorrectly reported data. It also does not use data that is not available in most countries or is only available for a few days. In addition, feature selection was implemented via forward selection for groups and then by backward selection inside the groups.

II. Data description and processing

Two data types are used: constant features and time series features. The dataset contains epidemiological, demographic, sanitary, continent, mobility, weather and policies. The data contains 34 features, 13 of which are constant for a particular country and 21 vary by day. Our analysis takes into account 115 countries with available data. We aim to model the reproduction number labeled as 'average r estim' each day based on these features.

III. Models and Methods

We employ various methods for choosing the most important features

- **Permutation feature selection**
  The idea of the method is straightforward: the importance of each feature is determined by how much the model prediction error increases after permuting this feature. The more the error increases compared to the baseline one, the more important this feature is for outcome since shuffling of features allows to break the dependence of the target variable on this feature.
  Baseline error was defined as an error of the model on test data for this model (which in our case means data for a certain country) and feature importance was defined as a difference between error obtained on test data with permuted feature and baseline error. Since the data contains features that are constant for a particular country, only variable features were used in this algorithm. For them, shuffling for a specific feature was performed over all days and all time steps.
  This algorithm was implemented as a two-step method-determining the feature importance for a single country and for a set of countries.
  This algorithm has significant advantage - it does not require retraining of the model, but only estimates of the prediction error, which makes it computationally inexpensive. Another advantage is that it takes into account interactions with other features and breaks them as well as for target variable. [2]
  The disadvantage of the algorithm is that due to random shuffling, the results can vary. To reduce this effect, shuffling must be repeated several times and results are averaged. In our implementation of the algorithm, shuffling is performed 100 times for each feature.
- **Principal component analysis**
  PCA is a linear dimensionality reduction technique which projects the input data on a lower dimensional space while minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving a Singular Value Decomposition (SVD) problem that generates eigenvectors (principal components) along with the corresponding eigenvalues which indicate the relative im-

portance of each component. This technique is widely used as a pre-processing step for data analysis and modelling, since its an unsupervised approach independent of the subsequent model and very easy to implement. To retain interpretability, we use PCA only on the time varying features in the input data. Since the different features vary on different scales, its important to normalize the data for them to be comparable. The principal components are subsequently ranked based on their respective eigenvalues. As a heuristic, we keep only select top eigenvectors that cross some threshold (ex. 90%) in explaining the variance of the data. One disadvantage of this method is that the principal components lie in a distinct orthogonal space which is different from the original feature space, that might make it difficult to understand the effect of the original features on model outcomes.

- **Random forest and Boruta**
  Random forest is an ideal model to capture a non linear relationship between input and output. It considers an ensemble of decision trees and considers the average of the output of all the trees. To implement random forest for the given data, we transform the problem to a supervised learning problem with the output being average r estimate and the input being all the corresponding feature vectors. Although the data is in a time series format and hence the assumption of Independent and identical distribution of each data point does not hold which the Random forest implicitly assumes, we can get a rough idea of feature importance by modelling it with Random forest. The feature importance by random forest can be observed through the Gini score of each variable, The higher the Gini score of a variable, the higher its importance. Random forest has a drawback and the result by it can vary every time it is run because of the inherent randomness. To remedy the problem, boruta algorithm employs shadow features which are a permutation of the input features and appends it to the original database, increasing the database to twice it's original size. This is repeated multiple times to decrease the randomness and hence improve the predictions and select the most important features. For fitting the random forest we considered 100 estimators and for boruta algorithm, the max iterations were considered 50
- **Gradient importance**
  This method is commonly used to determine the importance of features in predicting time series values. Its idea is to find out the contribution of each input value to the prediction, which is determined by the derivative of the prediction for each input - the gradient. By averaging the obtained values over time steps (over a time window), you can determine the importance of features on a certain day. By averaging the values obtained over the number of days, you can determine the time step wise feature importance.

IV. Results

- **Permutation feature selection**
  According to the results of the algorithm, we obtain the ranking of variable features for all countries included in the analysis (115 countries), which is summarized in the below Table. Based on the resulting ranking, we can say that the most important data for predicting the value are data of mobility and weather. The mean-squared error averaged across all 115 countries over the time of the epidemic was 0.2525. Further, in accordance with the obtained feature ranking, 10 features of the highest importance were taken from the variable ones. When using data containing 23 features (taking into account constant and selected variable ones), the error averaged across all 115 countries over the time of the epidemic became 0.1855 (which means an improvement of more than 26 %).
  Feature rankings were also obtained for each of the 115 countries, reflecting the different contribution of certain features for a particular country.

Table I
FEATURE IMPORTANCE RANKING FOR 115 COUNTRIES IN THE PREDICTION OF R VALUE

Variable feature	Mean importance	Std. dev.
perceived temperature	0.02575	0.05544
transit stations	0.01552	0.04673
retail and recreation	0.01002	0.03736
gathering size restrictions	0.00909	0.04079
grocery and pharmacy	0.00831	0.03309
parks	0.00766	0.02679
workplaces	0.00717	0.03458
stringency	0.00714	0.0416
maximum temperature	0.00453	0.02929
minimum temperature	0.00449	0.02714
workplace closing	0.00359	0.04554
home confinement orders	0.00301	0.02618
cancel public events	0.0022	0.03635
close public transport	0.00147	0.01935
weekday	0.00136	0.00653
humidity	0.00134	0.0227
school closing	0.00091	0.03738
international travel restrictions	-0.00095	0.0261
internal movement restrictions	-0.00116	0.02993
pressure	-0.00207	0.01624
residential	-0.00283	0.02386

- **Principal component analysis**
  Figure 1 shows the cumulative relative eigenvalues by the principal components in order of their importance. It is evident that as we add more and more principal components, the added importance decreases, with the last few components showing negligible improvement. This indicates the possible redundancy in the input data, which is expected since many of the features are correlated.
  We can gain further understanding by observing the

composition of principal components by the original features (Figure 2). The first component shows large contributions from all the policy and stringency features that indicates a correlation between them as expected. Similarly, weather features such as temperature, pressure and humidity appear in the second component. Thus, this composition may help us understand the distribution of input data among various features and their interdependence.

For studying the impact of PCA, we use a 90% threshold to select top 10 principal components for training. The following table shows the obtained errors on a few countries after training with the modified input features. It is evident that the accuracy of the model has not changed a lot with improvements in some countries and losses in others, whereas the size of feature space has reduced significantly ($\approx 67\%$ of the original size).

Table II
CHANGES IN THE MEAN SQUARED VALIDATION ERROR AFTER USING FEATURES SELECTED BY PCA

Country (iso code)	Baseline error (34 features)	PCA features error (23 features)	improvement
CHE	0.093785	0.065177	0.028
USA	0.123717	0.150307	-0.026
IND	0.134485	0.106738	0.028
GBR	0.061276	0.074523	-0.013
BRA	0.256413	0.315258	-0.059
ZAF	0.115017	0.135353	-0.020
		Average error	-0.016

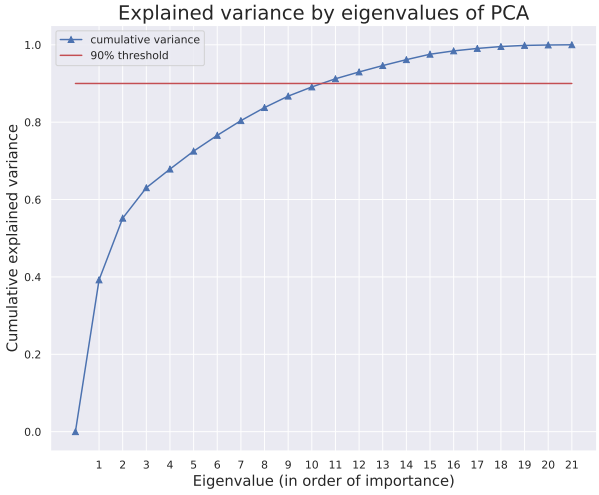

Figure 1. Cumulative relative importance of the principal components

- **Random forest and Boruta**
  From the random forest algorithm, we find that the most important feature is stringency followed by workplaces and life expectancy. The boruta algorithm on top of random forest classifies all the feature vectors as important and hence we can employ the feature importance by random forest to know the most important features. Figure 3) ranks the gini score of various features that were selected by the random forest. Higher the

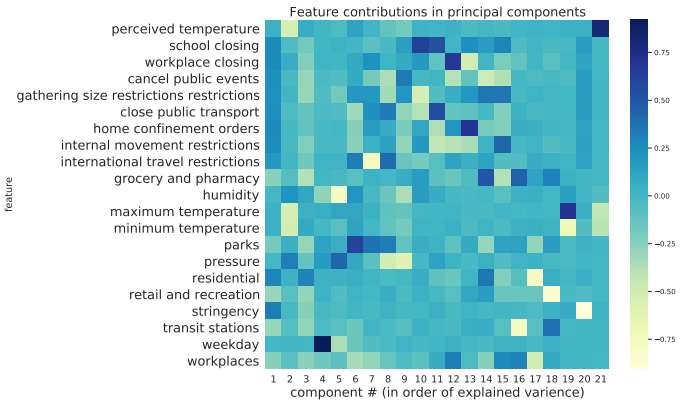

Figure 2. Composition of the principal components

gini score of a feature vector, more is it's importance according to the random forest. After fitting the random forest to the data, we get a mean absolute error of 0.33 by cross validation which is not as good as LSTM but is close to the error values obtained by the LSTM and hence can help us in choosing important features for prediction of Reproduction number.

Table III
CHANGES IN THE MEAN SQUARED VALIDATION ERROR AFTER USING 27 FEATURES SELECTED BY RANDOM FOREST

Country (iso code)	Baseline error	Random forest features error	Difference
CHE	0.0937855	0.121742	-0.0279565
USA	0.1237174	0.088943	0.0347744
IND	0.1344851	0.081524	0.0529611
GBR	0.0612765	0.048571	0.0127055
BRA	0.2564139	0.315545	-0.0591311
ZAF	0.1150178	0.131184	-0.0161662
		Average error	-0.00056256

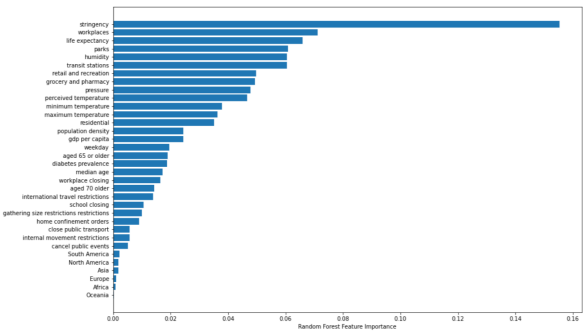

Figure 3. Feature importance by random forests

- **Gradient importance**
  By averaging the obtained gradient values over a time window, this algorithm provides information about which features are important on a particular day. This is reflected in Figure 4 and Figure 5 on the example of Switzerland for the 51st and 101st days of the epidemic. It can be noticed that for features gain significant importance on 101^{th} day such as transit stations, humidity,

grocery and pharmacy and some features loose importance such as international travel restrictions, parks etc. This information could be utilized to track the changes in feature importance over time and adjust the policies accordingly.

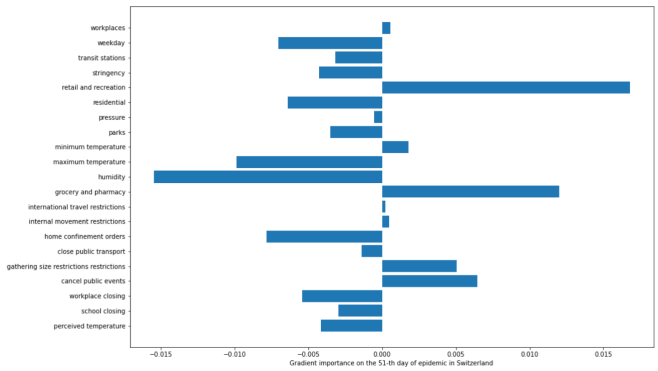

Figure 4. Gradient importance for Switzerland for 51st day of epidemic

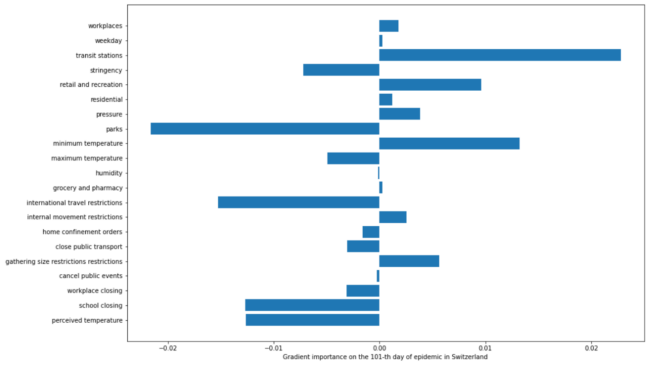

Figure 5. Gradient importance for Switzerland for 101th day of epidemic

V. Discussion

There is some similarity among most important features between different methods. According to random forest the most important features are stringency, workplaces, life expectancy, parks, humidity and transit stations and Boruta classifies all the features identified by the Random forest as important allowing us to directly use the gini score to compare the feature importance. By permutation importance we get the most important features as perceived temperature, transit stations, retail, gathering size restrictions, grocery and parks. 2 of the top 6 features are common between both methods. Parks and transit stations suggesting the positive effect of these features on the reproduction number.

Here are some criteria for comparison of the different algorithms implemented:

- **Interpretability**
  - Permutation feature selection - provide mean importance for each feature averaged over all countries. Feature importance for a certain feature for a given country was defined as a difference between error obtained on test data with permuted feature and baseline error.
  - PCA - since the outcome of PCA gives us principal components which are a linear combination of original features, it is non-trivial to understand importance of each individual feature independently, which makes this algorithm obscure in terms of interpretability
  - Random forest - Random forest outputs the feature importance by the gini score
- **Computation time**
  - Permutation feature selection - since this algorithm does not require retraining the model, it is relatively computationally cheap.
  - PCA - Since the algorithm is independent of the model used, it is computationally very cheap to implement. The time complexity typically scales as $\mathcal{O}(d^3)$ where d is the size of input feature space
  - Random forest is relatively cheap to run but Boruta algorithm is expensive to run as it permutes the feature vectors several times to find the most important features

VI. Limitations and Future Work

For feature selection, Random forest is not appropriate as it is fit on a time series data which is not independent and identically distributed and hence is not an ideal setup for a supervised learning algorithm. The importance of features we get through Random forest hence should be only taken indicatively and should be refined after cross checking with other methods. PCA is model independent and can only help in explaining variance of the data but not aid in creating a model which can predict values in future for different data sets. For the permutation feature importance method you can select a different number of the most important features according to the ranking so it is possible to vary amount of selected features and observe errors in order to minimize them. Further work is also required to analyze the results of the gradient importance method.

REFERENCES

[1] T. Bossy, *Identification of the Dynamic Drivers of Effective Policy during the COVID-19 Pandemic*. EPFL, 2020.

[2] Christoph. Permutation feature importance. [Online]. Available: https://christophm.github.io/interpretable-ml-book/feature-importance.html