Data Science for Social Good UK(DSSGx UK)

Case Study

Mapping the World's Offline Population (DSSGx UK 2021)

PROJECT PARTNERS

The International Telecommunication Union (ITU) is a specialised agency of the United Nations responsible for all matters related to Information and Communication Technologies (ICTs). ITU members include 193 UN member states as well as some 900 companies, universities, and international and regional organizations. ITU has a technical mandate to coordinate the global allocation of common goods like radio spectrum and satellite orbits and to develop and maintain the standards that networks and ICTs can interconnect. It also has a development mandate improve the access to ICTs of underserved communities worldwide and to protect and support the right to communicate. In the context of the latter mandate, ITU and joined forces in 2019 to create Giga: a global initiative to connect every school to the Internet by 2030. This partnership provided the original motivation for this project.

THE PROBLEM

In 2022, 66% of the world's population have access to the internet. Conversely 34% - 2.5 billion people - do not. This figure rises to 64% in the least developed countries. It is very difficult to gather accurate data on the location and circumstances of the world's offline population since, by definition, they do not leave digital footprints. Aside from directly benefitting children in terms of education, schools often act as community hubs and so connecting a school to the internet often connects the surrounding community too. The Giga project therefore produces a secondary benefit for ITU in reducing the world's adult offline population. It would be desirable to include these secondary benefits among the criteria that are considered when prioritising schools for connection under the Giga initiative. The difficulty is that the offline population is not visible in data except at a highly aggregated level of description, often at the level of national statistics. Such aggregated numbers are not helpful in making decisions at the scale of a school district. The challenge that the ITU brought to the 2021 DSSGx UK summer programme was to use machine learning to combine multiple geospatial, technical, economic and sociodemographic data sets to predict estimates of the proportion of the population without access to the internet at a local level.

DATA SCIENCE FOR SOCIAL GOOD (DSSG) PROJECT

The DSSGx UK team took up the challenge to build a machine learning model that can integrate multiple different data sources to predict real-time estimates of the number of offline people in local areas that are more current, granular, and accurate than the estimates that are available from aggregated statistics. The primary aim was to identify those areas that would benefit most from extending Internet connectivity with an initial focus on schools as points of interest to feed into the Giga initiative. A secondary aim was to use the model to obtain estimates of regional and national connectivity statistics for countries for which such data is missing or very out-of-date.

The data that were available as inputs included internet speedtest data (Ookla), mobile cell tower data (OpenCellId), population data (WorldPop), Facebook user data and satellite data (night lights, global human modification index, vegetation).

To build a prototype, it was decided to focus initially on Brazil. This was because Brazil was willing to provide survey data on household internet connectivity on an enumeration area level to provide a ground truth target variable on which to train and validate a machine learning model. School location data was provided by UNICEF. Like most survey data, the ground truth connectivity data was patchy, quite sparse and at a different level of granularity to the school data. Intersecting the two data sets left a training set with about 11,000 schools.

Using the MLFlow framework, most of the standard regression algorithms – linear regression, random forest, XGBoost, Light GBM, SVM and neural network - were trained, tuned, and assessed. The performance was judged based on predictions for schools with the lowest connectivity – less than 30% - since these are the ones of most interest to Giga. By this measure, the random forest and XGBoost models turned out to perform best. However, an initial attempt to directly apply the model developed for Brazil to Thailand indicated that the model did not generalise well to other countries. This illustrated that further work would be required to apply the methodology developed in the project more widely.

A presentation of the project from the 2021 online Datafest event can be viewed at

- https://youtu.be/wD87TfWG2ts

Documentation and sample model output can be found here:

- https://dssgxuk.github.io/itu/

FOLLOW UP POST PROGRAMME

After the end of the summer fellowship programme, ITU hired DSSG fellow Utku Can Ozturk as an intern to continue working on the project. The objective was to make the model pipeline more robust so that it could reliably be used to train models for other countries. It had become clear during the summer project that it was necessary to adopt a systematic approach to bringing all the input data to a common spatial resolution in a principled manner, and that restricting connectivity predictions to areas around schools was an unnecessary limitation for the use of the tool. Therefore as part of the follow-on work it was decided to reconfigure the pipeline to use a tile-based approach to estimating connectivity on a hexagonal grid covering an area of interest. Input data could then be standardised to a required resolution by setting the tile size and averaging. The code was refactored and streamlined to allow it to incorporate a wider variety of different data sources including satellite data on topography, population stratification by age and recent ITU datasets on optical fibre deployments. This made it much easier to run for multiple countries: it has now been tested against ground truth survey data from Brazil, Mexico, Georgia and Zimbabwe. Utku is now working full-time as an analyst for the ITU infrastructure data team.

The ultimate objective of the continuation project is to be able to provide ITU member countries with a sufficiently simple set of tools that they will be able to train their own model on their own survey data to get around the need to share potentially sensitive data with ITU. To achieve this, analysts from interested countries will require the provision of some training and the development of a user manual. This is to ensure that countries are able to train and validate their models correctly because the ability of the model to predict meaningful estimates of connectivity is highly dependent on the quality and

quantity of ground truth survey data available to train it. Discussions are under way with member countries to move in this direction but it hasn't happened yet.
The project has now been incorporated into an open data workstream within ITU and will become more visible both internally and externally over the coming months since it provides a tangible example of the kinds of benefits that can be obtained from open data projects. ITU are actively presenting and promoting this work project to other similar organisations. In June 2022, the ITU team presented the work to the World Telecommunication Development Conference in Kigali.

"The DSSG experience has been amazing – the fellows are embarrassingly smart and I'm impressed by the quality of their work and their dedication to the project."

Youlia Lozanova, Senior analyst – policy and regulation, ITU