Macro-Eyes Machine Learning Scientist Challenge

Thank you for your interest in Macro-Eyes and willingness to participate in our technical challenge!

The technical challenge emulates some of the work we do at Macro-Eyes. It has two parts:

1. a prediction problem, and
2. an open response problem.

For both, we are interested inseeing how you think through problems when confronted with ambiguous problems and messy data.

You have 72 hours from when you download the exercise to complete it. We estimate it will take around 5 hours. At the end, please send all code, notebooks, and written responses to:

- ML Scientist Upload Form

If you have any questions, you can also email Johannes, Johnna and Giacomo (johannes@Macro-Eyes.com, johnna.sundberg@Macro-Eyes.com, giacomo@Macro-Eyes.com) and we will get back to you as soon as possible (maximum 24 hours). If you have technical questions about the download or upload, please reach out to Beverly.Kehoe@macro-eyes.com.

PART 1: PREDICTING IMMUNIZATION DROP-OUTS

Context: Although vaccination rates have increased globally over the last twenty years — largely due to efforts to ensure vaccines are stocked at convenient points of care even in remote locations — they have plateaued in the last decade. This is largely attributable to children who drop out of their vaccination schedule, i.e., do not receive all their required vaccines, despite access. Delayed vaccination puts many children at risk and often requires costly vaccination campaigns to resolve.

Problem Statement: Imagine you are working with an organization that runs health clinics in Botswana. They want to be able to send health workers to follow-up with children who have not yet received all 4 doses of OPV and 3 doses of DPT at 4 months of life. They cannot individually follow-up with all children, so your job is to help them target their intervention by predicting which children will not become vaccinated by 6 months without intervention. Therefore, you can use all information; for example, which vaccines the child received, up until 4 months after the child is born.

Below are some examples of scenarios when health care workers would and would not intervene:

- Child A received 4 doses of OPV and 3 doses of the DPT vaccine by 4 months of life. Because this child received all the vaccinations, the health care worker would not intervene.
- Child B received 2 doses of OPV and 1 dose of DPT by 4 months of life. At 5.5 months, Child B received 1 more dose of OPV and 1 more dose of DPT, and then no more vaccinations until 8 months. Because the child was still undervaccinated at 6 months, the model should tell the health care worker would intervene at 4 months.
- Child C received 3 doses of OPV and 2 doses of DPT by 4 months of life, and then no further vaccinations. As Child C did not complete the vaccinations, a health care worker intervenes.

You may complete exercise using the language (Python/R/etc.) and tools of your choice.

Data Dictionary: patients_db.csv

- pat_id: The unique ID of the child.
- dob: The date of birth of the child.
- gender: The gender of the child.
- fac_id: The unique ID of the health facility the child received the vaccination.
- lat: The latitude of the facility.
- long: The longitude of the facility.

- district: The geographical district that the facility is located in.

Data Dictionary: immunizations_db.csv

- pat_id: The ID of the child.
- vaccine: The abbreviated name of the vaccine the child attempted to receive.
- im_date: Immunization date, i.e., the date the child received the vaccine.
- successful: Whether or not the vaccination was successful.
- reason_unsuccesful: If the vaccination was unsuccessful, the selected reason why.

Below is also the WHO infant immunization schedule for your reference:

- OPV (oral polio vaccine):
    - Dose 1: birth
    - Dose 2: 6 weeks
    - Dose 3: 10 weeks,
    - Dose 4: 14 weeks*
- DTP (diphtheria, tetanus, pertussis):
    - Dose 1: 6 weeks
    - Dose 2: 10 weeks
    - Dose 3: 14 weeks*

* Although the final doses of these vaccines are supposed to occur just after 3 months, health care workers often allow a grace period of a few weeks to allow children who were slightly off schedule to catch up before following up with children who are still unvaccinated. Therefore, we chose to use 4 months instead of 14 weeks as the initial cut-off.

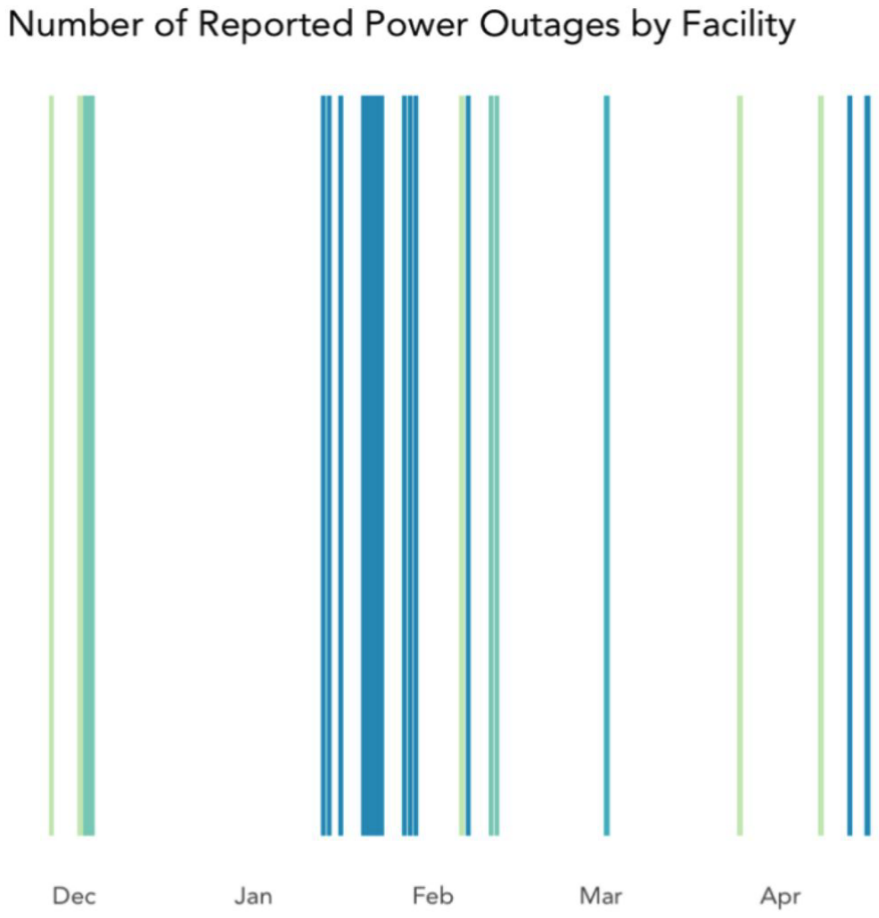
PART 1 DELIVERABLES:

- Predictions for which children will or not receive all vaccines by 6 months on a test set of data (generated by you)
- All notebook and scripts used to create predictions:
    - Data exploration
    - A defined baseline
    - Predictive model, can you beat the baseline?
- A description of next steps:
    - If you had more time to work on this problem, what other data sources or methods would you like to add that you think would improve the predictions, and why?
    - How can your results help health facilities allocate health workers?

PART 2: UNSTRUCTURED INFORMATION FROM SMS

Problem statement: In lower-income countries, there is little visibility into the health system. To combat this, Macro-Eyes began asking health care workers on the ground to send in their daily observations about whatever they think is notable. Our inspiration for this was from recent advances in using unstructured data to predict events, for example using news data to predict stock prices. Our ultimate goal is to be able to use the messages to improve the health supply chain. For example, we have used the messages to measure reports on compliance with COVID prevention. Other than the text messages, we get monthly consumption data from the clinics, such as how many children they were able to vaccinate by vaccine by month.

As an example, here is a sample graph we put together to allow health officials to understand in which facilities power outages occur more frequently:



Furthermore, here are some additional sample messages:

- Empty and flooded health facility due to the rain.
- Good morning, improvements in the road and transport today, very hot day and marked with the opening ceremonies of the school year. The parents are busy sending their children to school, so there is little flow in the health facility today.
- Good afternoon, the Extended Vaccination Program today is calm this afternoon but, in the morning, we had a high number of patients.

DELIVERABLE FOR PART 2

Imagine that you were the lead data scientist on this project. Please outline in 1-2 paragraphs how you would approach this project and what types of problem statements you would formulate in order to help improve the health supply chain.

We are trying to gauge your ability to turn an abstract problem (increasing visibility into the health system) into actionable steps.

SUBMISSION INSTRUCTIONS:

Please upload your submission to:

ML Scientist Upload Form