Predicting Dengue Outbreaks with Explainable Machine Learning

Robson Aleixo[1], Fabio Kon[1], Rudi Rocha[2],
Marcela Santos Camargo[2], **Raphael Y. de Camargo**[3]

INCT InterSCity
http://interscity.org

[1]Department of Computer Science, University of São Paulo, Brazil
[2]São Paulo School of Business Adm., Institute for Health Policy Studies (IEPS)
[3]Center for Mathematics, Computing and Cognition, Federal University of ABC

AI4Health, May, 2022

CAPES FAPESP CNPq

Introduction

- Dengue disease: Mosquito-borne tropical disease
- CDC estimates hundreds of millions infections
  - Tens of thousands deaths per year

- State of Rio de Janeiro, in Brazil
  - Located in a tropical area and most cities are frequently under high-risk state of alert

https://pt.wikipedia.org

- Prevention of Dengue outbreaks
  - Reduce the mosquito population
  - Work-intensive: requires agents to visit locations

source: ministry of health

Objectives

- Develop a Machine Learning model for predicting Dengue outbreaks up to 3 months in advance
  - Use of multiple features, such as environmental, epidemiological, demographic, and spatial
  - Provide explainable predictions for health agents
- Evaluate the model using 6 years of data from Rio de Janeiro metropolis

Related Work

- Extensive literature on predicting Dengue cases and correlate with climatic and socioeconomical variables
  - Most works on tropical countries: Thailand, Indonesia, Malaysia, and Latin America
  - Used different ML techniques, from Linear Models to Neural Networks

- A limitation of existing work is that they provide little explainability of predictions
  - Linear models provide limited interpretability
- A second limitation is the use of different datasets, which are unavailable, and metrics
  - We will make all model code and data publicly available
  - We evaluate and compare several regression and classification metrics

Data and Feature Extraction

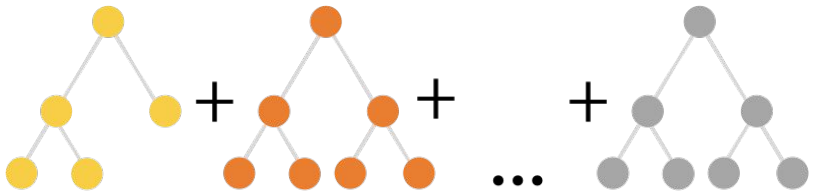
- Obtained from government agencies, such as
  - National Health Notification Information System (SINAN)
  - National Register of Health Facilities (CNES)
  - National Weather Institute (INMET)

- Dengue cases: last 3 months, prevalence, neighbors
- Environmental
- Mosquito infestation in the city and similar
- Demographic density and health coverage

Feature
cases m-1
cases m-2
cases m-3
dengue_prevalence
neighbor_cases
precipitation (mm)
temperature (°C)
air_humidity (%)
liraa
chikungunya
zika
demographic density
num_health_unit

Regression Method

- Boosted-tree regression method: CatBoost
  - Series of simple trees, where each tries to improve the prediction of previous trees
  - Captures non-linear relationships between features
- Baseline: Seasonal Autoregressive Integrated Moving Average (SARIMA)
- Used data from 2015 to optimized hyperparameters

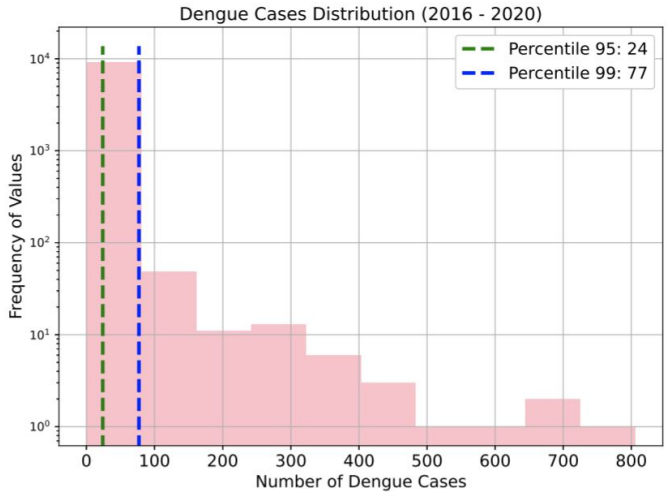


https://catboost.ai/

Predictions

- Predictions for each district, using data from 2016 to 2020
- CatBoost: 5-fold validation, using 4 years for training and 1 for testing
  - 3-month multistep-ahead predictions
- SARIMA model: separate model for each year
  - Used the four previous years to adjust the model

- Evaluated Regression and Classification errors
  - Regression: R^2, MAE, MAPE, RMSE
  - Classification: precision and recall
    - Severe and Mild outbreaks



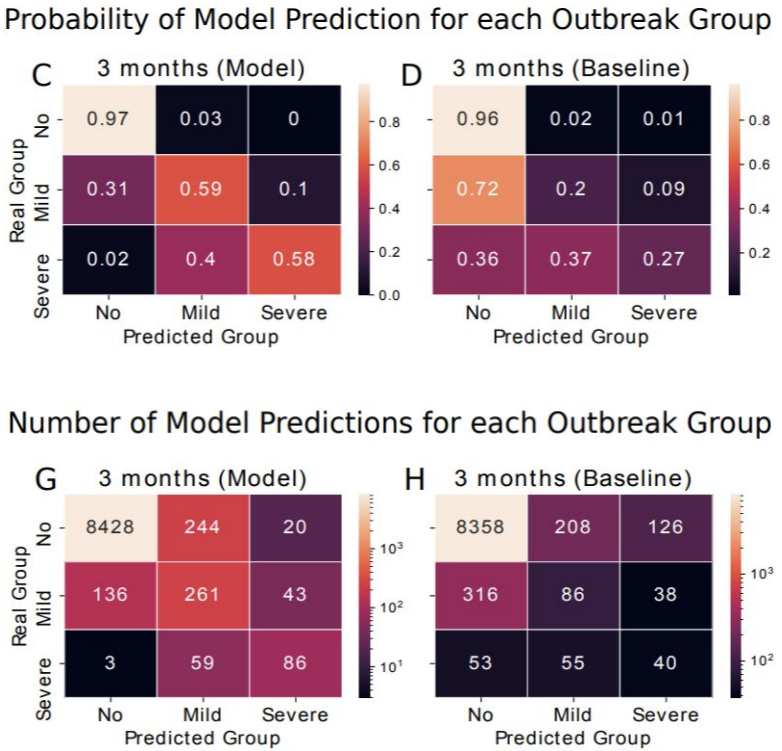
RESULTS

Outbreak Predictions 3 months in advance

- Three months in advance provides enough time for health authorities to act

- Predicted no outbreaks 97% of the time
- Recall of 59% (58%) for mild (severe) outbreaks
- When predicting an outbreak as mild
  - 57% chance of a mild or severe outbreak
- When predicted as severe
  - 86% chance of a mild or severe outbreak

- SARIMA model had more modest performance
  - Precision of 40% and 38% for mild and severe outbreak predictions

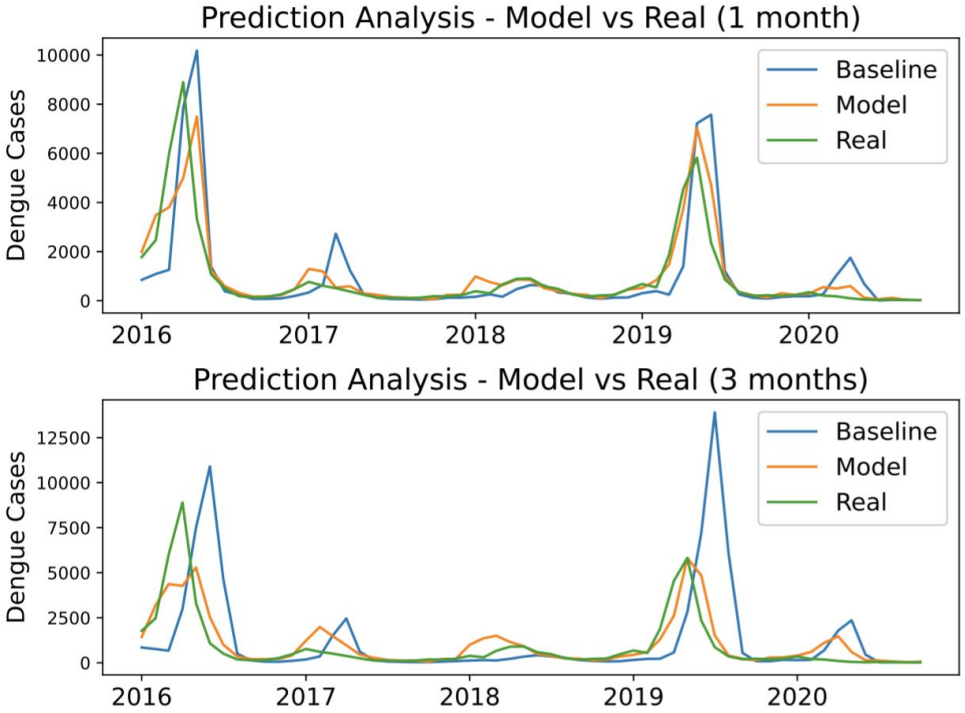


Seasonal Effects

- Dengue has a seasonal pattern
  - High-season (January to May)
  - Varying outbreak degrees

- Our model captures well the seasonality
  - Results for one month are closer to real values, but three-months also worked well
  - It tended to overestimate when there were fewer cases overall

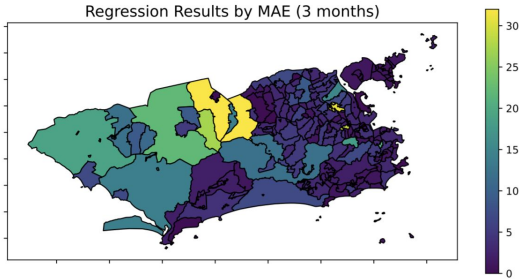
- SARIMA also captures the seasonality, although it tended to overshoot predictions

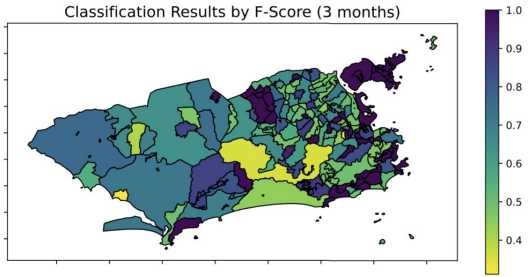


Evaluation per District

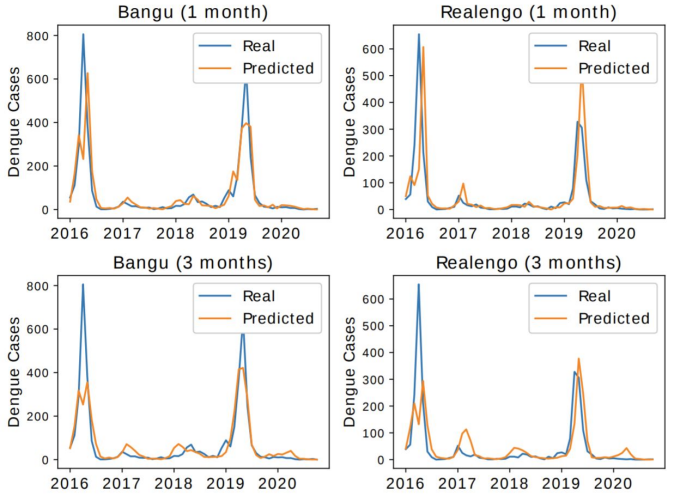
- MAE error unevenly distributed among districts
- Higher for districts which had very large number of cases, such as Bangu and Realengo
- F-score there were comparable to other districts, as the outbreaks were detected

- Harder to detect outbreaks in districts with fewer cases
- Remain mostly in the borderline of the outbreak threshold





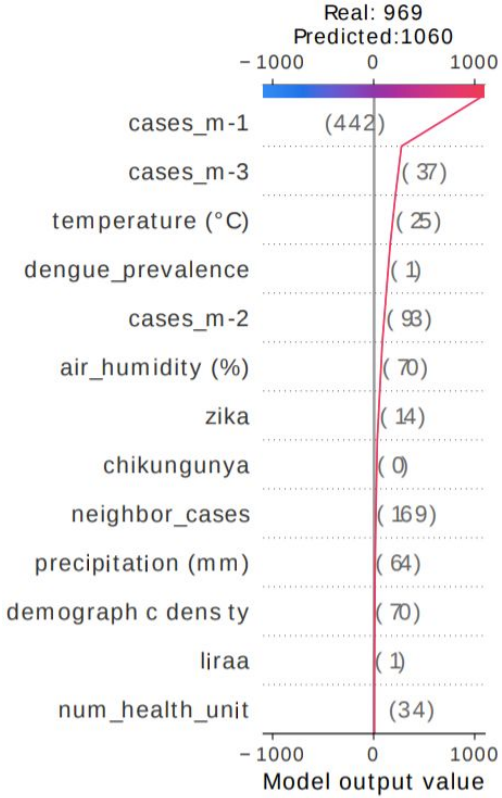
$$\text{F-score} = 2 \times \frac{precision \times recall}{precision + recall}$$

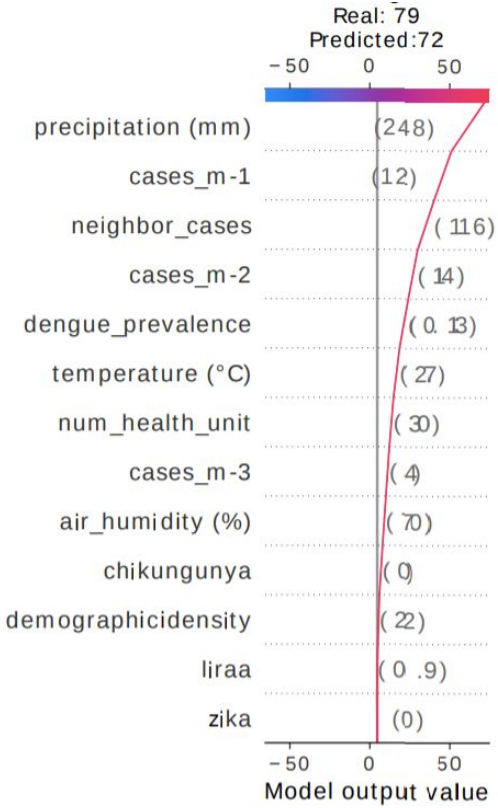


District	MAE	RMSE	MAPE	R2
Bangu	32.0	83.0	1.42	0.69
Realengo	32.0	81.0	2.62	0.42

Explaining the Results

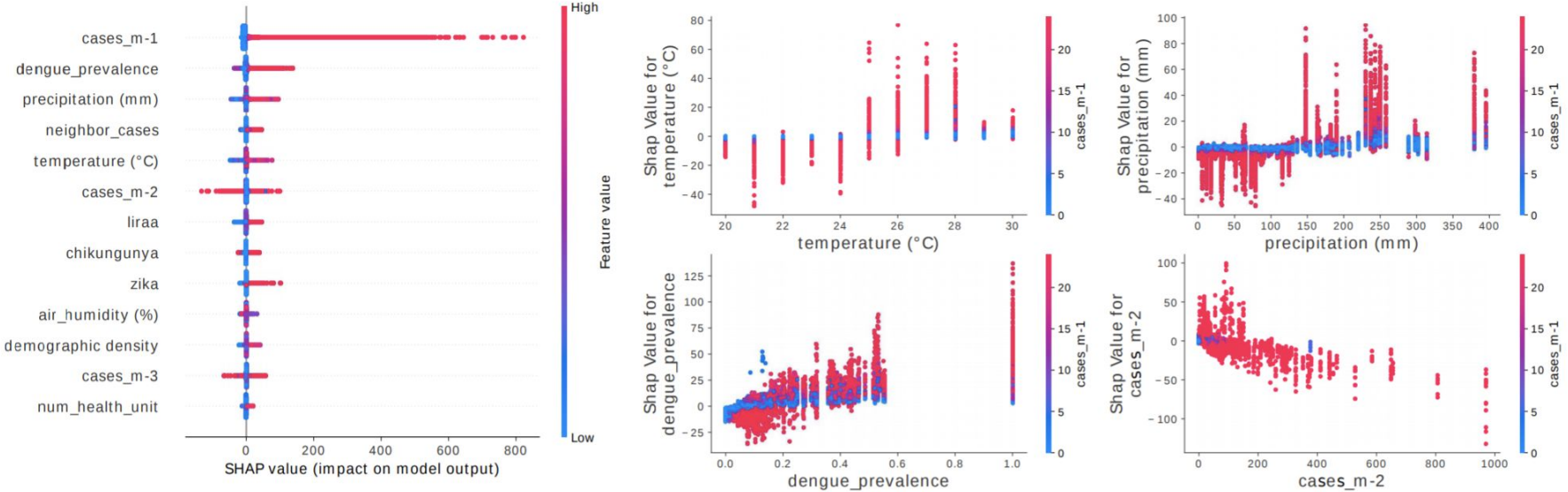
- SHAP (SHapley Additive exPlanations) values
  - Method to explain the contribution of each feature on each prediction
  - Has many desirable features: consistency, symmetry, additiviness, etc.





Explanations for the Complete Model

- Summary plot:
  - Features that most contributed to all predictions and the output direction
- Dependence plots:
  - Effects on predictions for feature values and dependence with another feature



Conclusions

- Developed an explainable machine learning model to predict Dengue outbreaks
  - Predictions three-month in advance provide health agents enough time to act
  - Combined with explanations for the predictions for better interpretation
  - Can be combined with other information available at health agencies

- Next steps
  - Enhancing the model with better features, such as sorotypes and other indicators
  - Extraction of better time series features and inclusion of data from more regions
  - Provide open-source visualization tool for health authorities

INCT
InterSCity

raphael.camargo@ufabc.edu.br

http://interscity.org

CAPES FAPESP CNPq