Problem description

Your goal is to predict the total\_cases label for each (city, year, weekofyear) in the test set. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively. You will make one submission that contains predictions for both cities. The data for each city have been concatenated along with a city column indicating the source: sj for San Juan and iq for Iquitos. The test set is a pure future hold-out, meaning the test data are sequential and non-overlapping with any of the training data. Throughout, missing values have been filled as NaNs.

* **Features**
* [List of features](https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#features_list)
* [Example of features](https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#features_eg)
* **Performance metric**
* [Mean absolute error](https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#mae)
* **Submission Format**
* [Format example](https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#sub_values)

The features in this dataset

You are provided the following set of information on a (year, weekofyear) timescale:

(Where appropriate, units are provided as a \_unit suffix on the feature name.)

City and date indicators

* city – City abbreviations: sj for San Juan and iq for Iquitos
* week\_start\_date – Date given in yyyy-mm-dd format

NOAA's GHCN [daily climate data](https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/" \t "_blank) weather station measurements

* station\_max\_temp\_c – Maximum temperature
* station\_min\_temp\_c – Minimum temperature
* station\_avg\_temp\_c – Average temperature
* station\_precip\_mm – Total precipitation
* station\_diur\_temp\_rng\_c – Diurnal temperature range

PERSIANN [satellite precipitation measurements](http://www.ncdc.noaa.gov/cdr/operationalcdrs.html) (0.25x0.25 degree scale)

* precipitation\_amt\_mm – Total precipitation

NOAA's NCEP [Climate Forecast System Reanalysis](http://rda.ucar.edu/datasets/ds093.0/#metadata/detailed.html?_do=y) measurements (0.5x0.5 degree scale)

* reanalysis\_sat\_precip\_amt\_mm – Total precipitation
* reanalysis\_dew\_point\_temp\_k – Mean dew point temperature
* reanalysis\_air\_temp\_k – Mean air temperature
* reanalysis\_relative\_humidity\_percent – Mean relative humidity
* reanalysis\_specific\_humidity\_g\_per\_kg – Mean specific humidity
* reanalysis\_precip\_amt\_kg\_per\_m2 – Total precipitation
* reanalysis\_max\_air\_temp\_k – Maximum air temperature
* reanalysis\_min\_air\_temp\_k – Minimum air temperature
* reanalysis\_avg\_temp\_k – Average air temperature
* reanalysis\_tdtr\_k – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's [CDR Normalized Difference Vegetation Index](https://www.ncdc.noaa.gov/cdr) (0.5x0.5 degree scale) measurements

* ndvi\_se – Pixel southeast of city centroid
* ndvi\_sw – Pixel southwest of city centroid
* ndvi\_ne – Pixel northeast of city centroid
* ndvi\_nw – Pixel northwest of city centroid

Feature data example

For example, a single row in the dataset, indexed by (city, year, weekofyear): (sj, 1994, 18), has these values:

|  |  |
| --- | --- |
| **week\_start\_date** | 1994-05-07 |
| **total\_cases** | 22 |
| **station\_max\_temp\_c** | 33.3 |
| **station\_avg\_temp\_c** | 27.7571428571 |
| **station\_precip\_mm** | 10.5 |
| **station\_min\_temp\_c** | 22.8 |
| **station\_diur\_temp\_rng\_c** | 7.7 |
| **precipitation\_amt\_mm** | 68.0 |
| **reanalysis\_sat\_precip\_amt\_mm** | 68.0 |
| **reanalysis\_dew\_point\_temp\_k** | 295.235714286 |
| **reanalysis\_air\_temp\_k** | 298.927142857 |
| **reanalysis\_relative\_humidity\_percent** | 80.3528571429 |
| **reanalysis\_specific\_humidity\_g\_per\_kg** | 16.6214285714 |
| **reanalysis\_precip\_amt\_kg\_per\_m2** | 14.1 |
| **reanalysis\_max\_air\_temp\_k** | 301.1 |
| **reanalysis\_min\_air\_temp\_k** | 297.0 |
| **reanalysis\_avg\_temp\_k** | 299.092857143 |
| **reanalysis\_tdtr\_k** | 2.67142857143 |
| **ndvi\_location\_1** | 0.1644143 |
| **ndvi\_location\_2** | 0.0652 |
| **ndvi\_location\_3** | 0.1321429 |
| **ndvi\_location\_4** | 0.08175 |

Performance metric

Performance is evaluated according to the [mean absolute error](https://en.wikipedia.org/wiki/Mean_absolute_error).

Submission format

The format for the submission file is simply the (city, year, weekofyear) and the predicted total\_cases for San Juan or Iquitos (for an example, see SubmissionFormat.csv on the data download page). The total\_cases should be represented as integer values.