

DengAI: Predicting Disease Spread

HOSTED BY DRIVENDATA

[HOME \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/80/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/80/)

[ABOUT \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/81/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/81/)

[PROBLEM DESCRIPTION \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/82/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/PAGE/82/)



[LEADERBOARD \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/LEADERBOARD/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/LEADERBOARD/)

[DATA DOWNLOAD \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/DATA/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/DATA/)

[SUBMISSIONS \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/SUBMISSIONS/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/SUBMISSIONS/)

[TEAM \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/TEAM/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/TEAM/)

[DISCUSSION \(HTTPS://COMMUNITY.DRIVENDATA.ORG/C/DENGUE-COMPETITION\)](https://community.drivendata.org/c/dengue-competition)

14

[OFFICIAL RULES \(/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/RULES/\)](/COMPETITIONS/44/DENGAI-PREDICTING-DISEASE-SPREAD/RULES/)

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 `NaN`s.

Features

List of
features
Example of
features

Performance

metric
Mean absolute
error

Submission

Format
Format
example

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/>)
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/detail.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.

For example, if you just predicted that there were 5 cases each week for 5 weeks in San Juan and 3 cases each week for 5 weeks in Iquitos, for a total of 10 weeks, you would have the following predictions:

city	year	weekofyear	total_cases
sj	2008	18	5
sj	2008	19	5
sj	2008	20	5
sj	2008	21	5
sj	2008	22	5
...			
iq	2013	22	3
iq	2013	23	3
iq	2013	24	3
iq	2013	25	3
iq	2013	26	3

Your .csv file that you submit would look like:

```
city,year,weekofyear,total_cases
sj,2008,18,5
sj,2008,19,5
sj,2008,20,5
sj,2008,21,5
sj,2008,22,5
...
iq,2013,22,3
iq,2013,23,3
iq,2013,24,3
iq,2013,25,3
iq,2013,26,3
```

Keep in mind that you need to submit one csv with predictions for both cities! Hence the requirement of the `city` column. Results will be parsed on our end and MAE scores will be given for each city's predictions.

Good luck!

Good luck and enjoy this problem! If you have any questions you can always visit the [user forum](http://community.drivendata.org/) (<http://community.drivendata.org/>)!

ABOUT DRIVENDATA

[What we do \(/about/\)](/about/)

[Who we are \(http://drivendata.co/#team\)](http://drivendata.co/#team)

[Blog \(http://blog.drivendata.org\)](http://blog.drivendata.org)

LEGAL

[Terms of Use \(/termsofuse/\)](/termsofuse/)

[Copyright Policy \(/copyrightpolicy/\)](/copyrightpolicy/)

[Privacy Policy \(/privacypolicy/\)](/privacypolicy/)

WORK WITH US

[As a partner \(/partners/\)](/partners/)

[As a competitor \(/competitors/\)](/competitors/)

[Join a competition \(/competitions/\)](/competitions/)

CONTACT

[info@drivendata.org \(mailto:info@drivendata.org\)](mailto:info@drivendata.org)

DrivenData Inc.

1062 Delaware St.

Denver, CO 80204