Challenge #2: Weather challenge

Context

Extreme weather events cause high material losses e.g. damage done by floods in 2010 in Poland is evaluated at around 13 000 000 000 PLN [1]. Because of that weather prediction and alerting is an extremely important task. Unfortunately, it is also a difficult one.

Currently, experts at IMGW produce alerts for dangerous weather manually by examining available forecasts, numerical models, satellite images and many other data sources.

Your task will be to help them automate part of the alerting system.

[1] - http://klimat.imgw.pl/wp-content/uploads/2013/01/tom3.pdf (Page 20)

Challenge

In this competition, your goal is to predict whether cumulative rainfall in the next 24h time frame will exceed 15mm in Olsztyn County with Olsztyn included (https://en.wikipedia.org/wiki/Olsztyn_County, https://en.wikipedia.org/wiki/Olsztyn). Prediction is valid if sum of precipitation recorded on **any** of the weather stations in this territory in the upcoming 24h will exceed 15mm threshold. You will do this prediction every hour.

This is directly related to an alert for intense rainfall produced by IMGW. It just has a different threshold for triggering and you don't have to provide the alert with time buffor before the weather event. Original IMGW alert is emitted when accumulated precipitation in 24h time frame is over 30 mm (source: http://www.pogodynka.pl/ostrzezenia/klasyfikacja):

Time span of data for this task is 01 Jun 2019 - 31 Aug 2019.

Dataset

Core dataset for this challenge are .csv files with various weather parameters measured in weather stations located in Olsztyn County (with Olsztyn included). Frequency of the data points depends heavily on the parameter, but all of them are placed in 01 Jun 2019 - 31 Aug 2019 time frame. There is one .csv file for each month and each parameter. Every row contains id of the source station, parameter code, timestamp in UTC and the value of the measured parameter.

We've also prepared some basic dataset of numerical weather predictions - ALARO model - for this time frame. It is not obligatory to use it, but it might be helpful. Weather predictions were cut beforehand to only first 30 (0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30) forecast hours to reduce amount of data. Nothing else was done to the original .grib files.

For more info please refer to the README files published along with the 2 above datasets.

Core dataset - train:
https://mega.nz/#F!Ilg0jQQb!-0b6ICdhJoRLezTb_zh1Dg

Core dataset - test (only .zip files as in train):
https://mega.nz/#F!sNgQyKgQ!QkI24LIiBIlacKLUUZXMqQ

Forecasts:
https://mega.nz/#F!xJQWzSQK!s30IokGacDrOEH8txR6NDA

Example external datasets

1. GFS (Global Forecast System):
   a. Data: https://nomads.ncdc.noaa.gov/data/gfs4/ (Focus on short forecast hours e.g. 0 - 24)
   b. Main site: https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs
2. IMGW public datastore of both historic and current data. Link: https://dane.imgw.pl/datastore

Evaluation

Your solutions will be evaluated on a separate test set cut out from the original time frame of 01 Jun - 31 Aug. Total volume of the test set is 3 separate, continuous weeks. They will be missing from the training data.

Metric used to evaluate your solutions will be standard F1 score derived from your binary prediction results on the test set.

CAUTION: As it is a time series prediction task it is very easy to cheat the final test. Team which will attempt this may be banned from future AiGames and will be excluded from evaluation.

Resources

1. PyGrib - very popular format for weather data. Usage: https://jswhit.github.io/pygrib/docs/
2. H5Py - reading .h5 format with Python. Link: https://www.h5py.org/