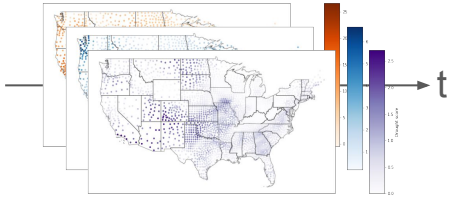
Machine Learning & Climate Project - Droughts forecasting

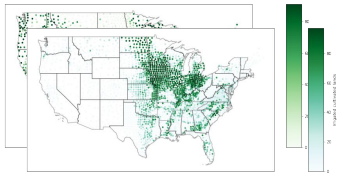
I. Context

- Climate change increases the odds of more frequent, intense and longer droughts
- Droughts impact agriculture, water supply availability, wildfires, etc.
- **Better forecast allows better resilience**

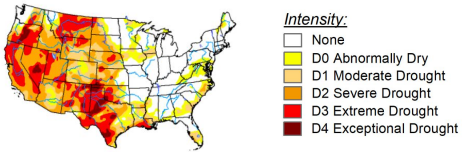
II. Data



Weather data
(daily observations of temperatures, precipitations, humidity, …)
18 variables



Soil data
(elevation, slope, aspect, land use, …)
30 variables

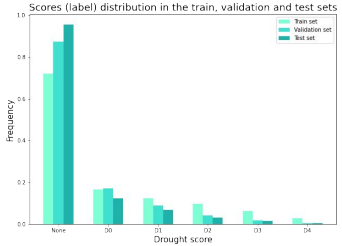


Observed droughts
(weekly observations)

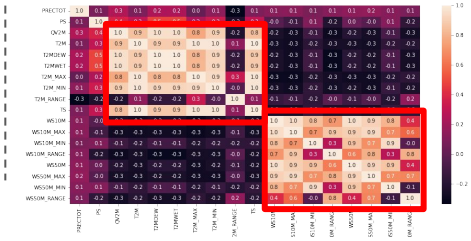
III. Objective

- Forecast droughts for **6** upcoming **weeks**
- **Regression** task on the continuous score
- Metrics used:
  - MAE / RMSE
  - **Macro F1 score** (binning predictions and targets)

IV. Exploratory Data Analysis



Highly imbalanced dataset
(rationale to focus on Macro F1 score)



Presence of **highly correlated variables**
(Temperature, Wind and Soil Quality variables)

Machine Learning & Climate Project - Droughts forecasting

V. Models

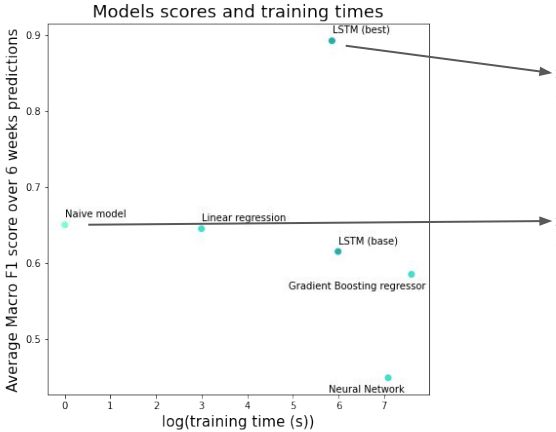
Naive model (baseline): Identity

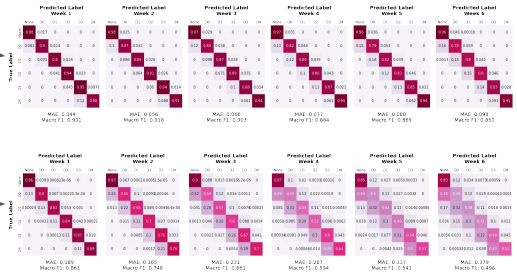
Non temporal models: Linear regressions, Gradient Boosting regressor, FFNN

Temporal models: LSTM

- Time Series are split into **subsequent episodes of a given length** to avoid correlation
- Matrix of observations is **flattened** or **averaged** for non temporal models

VI. Results





Confusion matrices of 6 weeks predictions for the **best LSTM** (above) and the **naive model** (below)

VII. Discussion

- Naive model's score opens a **discussion on the prediction's scope**
- Flattening the data causes an **explosion of the number of features** and suppresses all time depencies. Averaging data does not help

VIII. Eventual next steps

- Transformers
- Include weather forecast
- Best model predictions beyond 6 weeks