blueground
AI-Powered Groundwater Management

Water Shortage Prediction

Hi!ckathon #5

Organized by Hi! Paris

Team:

Florian Morel, Lucien Perdrix, Yaël Gossec,
Emma de Charry, Ahmed Yassine Chraa, Matteo Denis

Overview

Groundwater, often called the invisible lifeline of our planet, is facing an unprecedented crisis. Overuse, contamination, and the growing impacts of climate change are draining this vital resource faster than nature can replenish it. This silent depletion doesn't just threaten water availability—it imperils ecosystems, agriculture, and the very foundations of human resilience. For future generations, the stakes are monumental: a world where water scarcity disrupts livelihoods, diminishes food security, and destabilizes entire regions.

As a start-up at the forefront of sustainable water management, **BlueGround** plays a vital role in addressing the groundwater crisis. Collaborating within a European project of diverse actors, we harness artificial intelligence to develop predictive models that forecast water shortages and assess groundwater health. Our innovative solutions empower stakeholders to make informed decisions, ensuring the sustainable use of this critical resource for future generations.

Business Approach

We are a dynamic start-up dedicated to developing cutting-edge technological solutions to promote sustainable water management. As part of a collaborative European project uniting a diverse array of actors—including research institutions, industry leaders, and public organizations—we are at the forefront of efforts to tackle one of the most pressing challenges of our time: water scarcity.

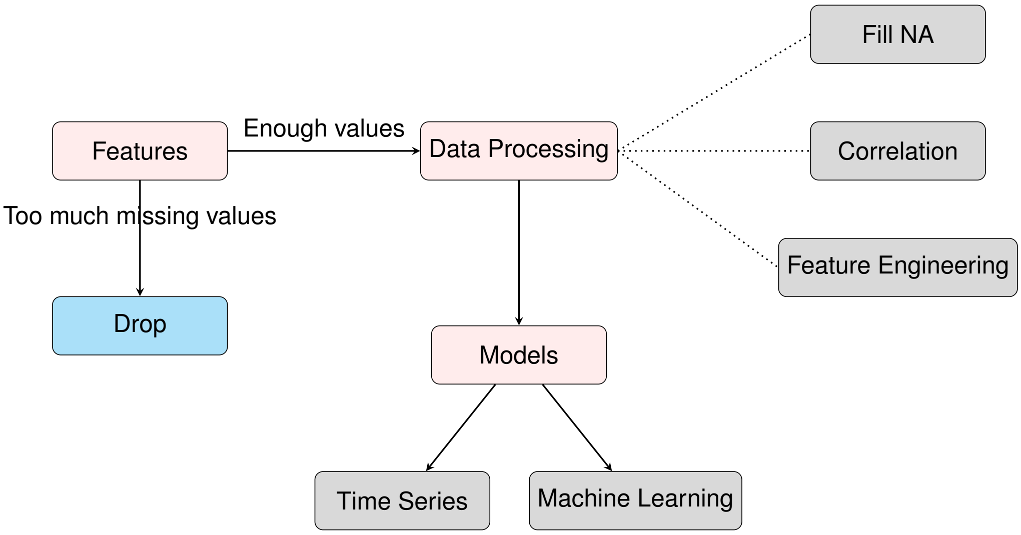
Our focus lies in harnessing the power of artificial intelligence to create predictive models that provide valuable insights into water shortages and the overall state of groundwater systems. By combining data from various sources and leveraging advanced algorithms, our solutions aim to support decision-makers in optimizing water use, preserving ecosystems, and ensuring long-term availability for communities and industries alike.

Through innovation, collaboration, and a shared commitment to sustainability, we strive to make a meaningful impact on water resource management across Europe and beyond.

Scientific Approach

Our strategy started with constructing a pipeline to train an initial model using a limited set of features. This pipeline was designed as a sequence of modular classes, each responsible for transforming specific groups of features. Once the baseline model was established, we progressively added more features and tested alternative models to enhance performance and refine predictions. This structured approach ensured flexibility and scalability in handling diverse feature sets.

To provide a clear understanding of our approach, we will first illustrate our method through a concise figure. This visual representation will highlight the key steps and relationships in our process, offering an accessible overview before diving into detailed explanations.



Our initial experimentation began with a K-Nearest Neighbors (KNN) model, which achieved a F1-score of 33%. Following this, we enriched our feature set by adding cleaned and processed features. With this improved dataset, we retried the KNN model and also tested Decision Tree and Random Forest models. The latter two significantly outperformed the initial results, demonstrating the impact of both feature engineering and model selection on improving predictive performance.

After testing various models, we ultimately chose a **Gradient Boosting Model** as our final approach. Its superior accuracy and robustness in handling complex relationships within the data made it the best fit for our task. With this model providing a strong foundation, we shifted our focus to refining and expanding the feature set.

About the features

Feature selection was key in our approach. A good model is first of all a model with good features. Hence, it was a key task to drop, transform, combine features from the initial dataset, to create a new base, on which our model would be able to learn properly.

In the initial dataset, features are dispatched in 5 section:

- Watertable
- Weather
- Hydrology
- Withdrawal Economic data

For this need, and to have a great synergy between the different members of the group, we adopted a **Pipeline** structure, which would allow us to develop different **Transformers** (= functions) and stack them all, treating sequentially the data, in different places. We were then able to work in parallel over different features, which made us gain loads of time.

Here is a short presentation of these pre-processing Transformers:

1. `DropNARate(rate)`: drops every column having more than *rate*% of missing values.

2. `TimeTNX()`: processes data relative to time maximum and minimum temperature at shelter during a day (= convert to float, fill missing values, ...).
3. `PrelevVol()`: processes data relative to withdrawal volumes.
4. `Prelev(usage_max_categories, mode_max_categories, scale)`: One-Hot encodes other features relative to withdrawal.
5. `CleanINSEE()`: processes data relative to economics.
6. `AltitudeTrans()`: processes data relative to altitudes.
7. `CleanLatLong()`: processes data relative to geographical position. Keeps only one position per row and transform others, especially the distance to nearest weather station.
8. `CleanTemp()`: Processes data according to temperatures. Impute missing values with a linear regression. Keep only four temperature features in the end.
9. `TemperaturePressionTrans()`: for temperature and pression features, if value is still missing, replace with the mean on the same day, in the same department.
10. `CleanHydro()`: processes data relative to hydrology.
11. `CleanPizo()`: processes data relative to watertable. Also One-Hot encodes some features.
12. (DateTransformer()): processes dates, keeping only one feature. Apply a cosinus function to have a view on the cyclicality of the dataset.
13. `DropCols()`: drop remaining features that we don't want to use in our model.
14. `StandardScaler()`: standardizes all quantitative features.

These transformers, all put together represent most of the pre-processing step. Once the pipeline is completed, we run a **Principal Component Analysis**, in order to have an idea of how many components to include. We observe, on the following figure, that from 30 components, the explained variance is not growing anymore, meaning that giving more features would not give more information to our model.

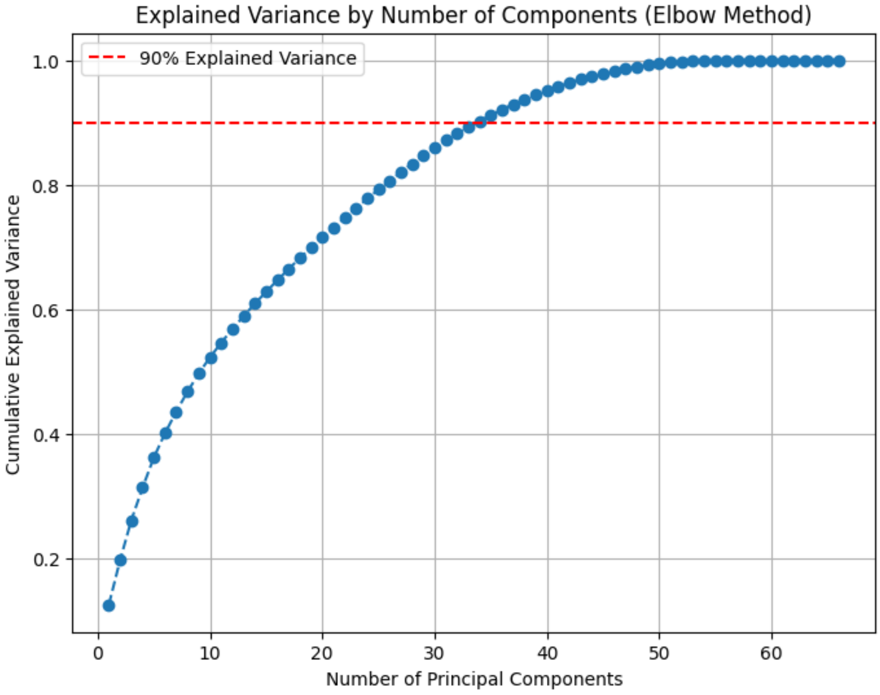


Figure 1: Principal Component Analysis

After the pre-processing and dimensionality-reduction steps describes above are done, there only remains to fit and run our chosen model, with the selected number of features (=components).

Results and Future Potential

With our approach, combining advanced feature engineering and a robust Gradient Boosting Model, we achieved a F1 score of X% in our predictions. This breakthrough offers a transformative solution for groundwater management, enabling precise forecasting of water shortages and providing actionable insights to optimize resource use, protect ecosystems, and ensure sustainable water availability.

As a start-up, **BlueGround** is positioned to become a key player in the field of water resource management. By leveraging cutting-edge artificial intelligence and collaborating within a European network of experts, we deliver innovative tools that empower stakeholders to make informed, forward-looking decisions. Our solutions have the potential to redefine how water is managed, making BlueGround an essential partner in addressing one of the most critical challenges of our time.