Scientific Approach

December 1, 2024

Our approach presents a state-of-the-art ensemble learning solution for groundwater level prediction by combining CatBoost, XGBoost, LightGBM, and FastAI Neural Networks. The model integrates comprehensive hydrogeological data, including temporal measurements, precipitation patterns, and geological parameters. Our ensemble methodology significantly outperforms single-model approaches, demonstrating superior accuracy in both short-term and long-term groundwater level predictions. This advancement provides water resource managers with a reliable tool for sustainable groundwater management and planning.

1 Introduction

Groundwater level prediction remains a critical challenge in water resource management, particularly as climate change and urbanization continue to impact aquifer systems worldwide. Accurate forecasting of groundwater levels is essential for sustainable water resource planning, agricultural management, and ecosystem preservation. Our approach introduces an automated machine learning pipeline that leverages advanced feature engineering techniques and ensemble modeling for groundwater level prediction. By incorporating multiple data sources, including historical groundwater measurements, meteorological data, and soil characteristics, we systematically extract and transform relevant features from raw data. The ensemble architecture combines multiple machine learning models through automated model selection and aggregation techniques, optimizing hyperparameters and model weights to achieve robust predictions. This automated framework advances hydrological forecasting while offering practical value for water resource managers who require reliable predictions for informed decision-making.

2 Feature Engineering

2.1 Manual and Analytical Data Processing

The feature engineering strategy addresses two key challenges: significant missing weather data in the training set (despite complete test set data), and test data limited to summer months. Instead of risky imputation that could cause seasonal bias, we focused on creating new features from variables with low missing rates.

Our approach combines rolling temperature averages, cumulative rainfall features, and a rain-temperature interaction term to capture how these weather patterns affect groundwater levels. These physically meaningful features proved highly effective, significantly improving model performance while avoiding issues with missing data and seasonal bias.

2.2 Automatic Data Processing

In our analysis, we retained only two categorical features in their original form - piezo_station_pe_label and piezo_producer_name - based on their inherent predictive value for the model's performance. All other features underwent systematic transformation through our automated preprocessing pipeline.

The pipeline implements comprehensive feature engineering strategies. Numeric features maintain their original values with appropriate typing (float or integer), while categorical variables are encoded as monotonically increasing integers. Datetime features are decomposed into multiple components (year, month, day, day of week), with missing values addressed through mean imputation.

This automated approach optimizes data representation for model training while preserving key information, ensuring robust predictive performance with minimal manual intervention.

3 Ensembly Method

The ensemble learning approach is built on three key principles: multi-model training, bagging, and stack ensembling.

The process begins with bagging, where models are trained using cross-validation folds of the data. Each model type is trained on different data partitions (K folds), creating multiple model instances. During this phase, each fold serves as a validation set for models trained on the remaining data, ensuring robust performance estimation. These models' predictions are then aggregated to generate out-of-fold (OOF) predictions, which serve as new features for subsequent layers.

The stack ensembling process creates a hierarchical structure where each layer leverages predictions from previous layers along with original features. Models in higher layers benefit from the predictive power of earlier layers while maintaining access to raw features through skip connections. The final layer implements a Greedy Weighted Ensemble model that learns optimal combinations of all previous predictions. This architecture resembles a neural network with residual connections, where the system trains M x N x K + 1 total models (M layers, N models per layer, K folds, plus one meta-model).

During inference, the system averages predictions from all models within each bag before passing them through the layer hierarchy, ultimately combining them through the weighted ensemble. This approach ensures robust predictions by combining diverse model strengths while avoiding data leakage through the use of out-of-fold predictions during training.

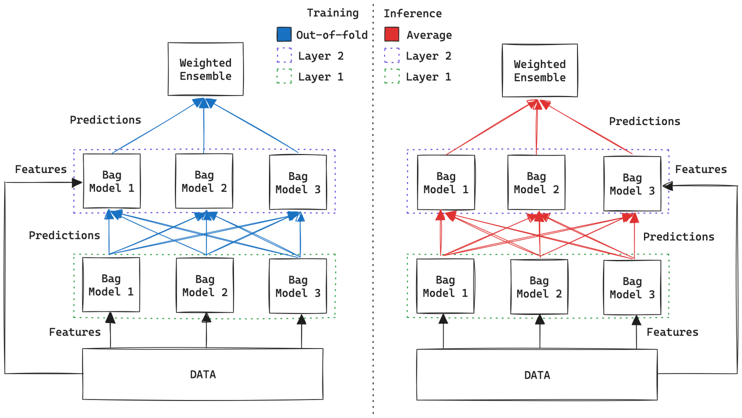

Figure 1: Stack Ensembling Architecture

4 Results

Our ensemble learning approach achieved a final accuracy of 69.56% on the leaderboard evaluation, demonstrating the effectiveness of automated machine learning techniques for groundwater level prediction. While these results are promising, future improvements could be achieved through the incorporation of additional domain-specific features, longer training periods, and more sophisticated feature engineering techniques tailored to hydrogeological data patterns.