PREDICTIONEER REPORT

Raj Singh Yadav 22B0341
Varun Ram Narayanan 22B0347

January 25, 2024

Contents

1 Overview

- We have created a model using various concepts of machine learning that takes into account various factors like rainfall, temperature, and water levels in order to predict depth of groundwater in different water bodies.
- We have carried out extensive Exploratory Data Analysis to fill the missing values spanning over 50% of the dataset.
- Then we applied our domain knowledge and also analyzed the dataset to determine useful features(Feature Selection and Feature Engineering).
- We have the data for the target variables for the next 3 years ranging from **29-06-2010 to 28-06-2013**.
- Following this, we tested various ML Models until we arrived at one to achieve the ideal results we wanted.

2 Data Preprocessing and Visualization

- We used standard functions available in the libraries pandas and numpy to get an idea of what data we have available at the time. We needed to fill the missing values now.

- We performed univariate feature analysis making use of matplotlib and seaborn to visualize the distribution of data over the years.
- We dropped the rows corresponding to the years 1998 to 2006 given the fact that almost all values were missing as a result of which there was no valid basis to predict the depth on.
- We dropped one depth column,Depth_to_Groundwater_DIEC in which there was no data.
- Two of the volume columns were dropped as they were composed of only zeroes and NaN values so their correlation with the target variable was undefined.
- Our target variables which we want to predict are now : **Depth_to_Groundwater_LT2, Depth_to_Groundwater_SAL, and Depth_to_Groundwater_CoS**.

3 Exploratory Data Analysis

- We plotted the distribution of all the features and target variables against the date column to understand their distribution with time.
- We plotted a heatmap in order to understand the correlation of feature variables with the target. We combined the insights from this with domain knowledge to correctly identify the most important features.
- The key observation we made in our analysis was that **the target variable data was more correlated with volume and hydrometry as compared to rainfall**.
- The volume of water in the system and the hydrometry are more **immediate indicators of water availability**. These factors take into account not only the current precipitation but also the accumulated effects over time, considering factors like water storage, retention, and flow dynamics. Rainfall is seasonal in nature as well which can cause a delayed impact on the target variable.
- We **smoothened** the features and cleaned the data to the maximum extent possible for the highly correlated columns.
- We used a **RandomForestRegressor** to fill in the missing null values. We also tried other methods like a **KNNImputer and an IterativeImputer** which gave very similar results while cleaning data.
- We plotted the data once again after this to analyze the effect of the imputer models.
- The data which replaced the NaN and null values seemed to be acceptable.

4 Model Creation

- In order to ensure no overfitting or underfitting in the model, we first scaled the data in all the columns using a **StandardScaler** as we decided to use an LSTM Model.
- We divided the model into a training and testing set to evaluate it.
- We observed the data for LT2 to be corresponding to an confined aquifer while the data for the other 2 regions SAL, and CoS was corresponding to the unconfined aquifer.Based on this we discovered the dependence of the target variables on the features.
- We had to predict data spanning the next 3 years. As we did have any feature data, we adopted an **LSTM (Long Short-Term Memory) model.**
- We used a sliding window approach to create sequences of input-output pairs for training the LSTM model.
- The model was trained using the prepared sequences of input-output pairs. During training, the model learned to capture temporal patterns and dependencies in the groundwater depth data.
- The model then generated predictions for the future time steps, providing an estimate of groundwater depth over the next 3 years based on the learned patterns from the training data.
- We made use of one hidden layer in the LSTM, along with activation function **relu**, and an **adam** optimizer.
- The main aim of the model was to minimize MSE which is a good measure regarding the accuracy of our model.