Udacity ML Engineer Nanodegree
Capstone Project
Acea Smart Water Analysis

Rose Koopman

1 Domain background

In the last decade humanity faced unprecedented challenges due to rising sea water level, flooding of rivers, draughts and overall more extreme whether conditions. With the effects of climate change becoming a severe thread to humanity, the monitoring of water supplies becomes of increasing importance.

With a service area covering over 9 million people, the Acea Group [1] is one of Italy's leading companies in the field of water supply. Its main responsibilities include building and maintaining a water network and preserving the health of its water bodies, in order to guarantee the daily water supply to Italy's inhabitants.

The Acea Groups utilises different types of water bodies:

- Aquifer: A water-containing layer (such as sand) in the ground. Water can be taken from the aquifer by making a well.
- Water spring: The place where an underground waterbody (e.g. underground river) emerges at the earth's surface.
- River: A flowing body of water, typically flowing from a water spring to the ocean.
- Lake: A contained body of water which is surrounded by land. There can be a river feeding into the lake and another river draining the lake.

2 Problem Statement

In order to provide in the daily water consumption, Acea Group carefully monitors the water level and water flow in its different water bodies. Each of these types of water bodies have their own characteristics and respond in a different way to events of rainfall or periods with high temperatures. Typically water bodies fill in autumn and winter when more water enters the bodies then is being utilised, while in spring and autumn water levels drop. In order to guarantee continuous water supply and preserve the health of the water bodies, there is a need to forecast water availability in terms of level and flow.

3 Datasets and inputs

The datasets are obtained from the Kaggle competition *Acea Smart Water Analytics* [2]. Nine different datasets are provided, corresponding to nine different water bodies. As the water level and water flow can be measured as several geographically located points around the water body, there can be multiple target values per water body. The provided features relate to weather conditions such as rainfall, and characteristics of the water body such as volume. Both the type of target value (level or flow) and the features differ per type of water body due to the different nature of the water bodies. Data is available on a daily level for some water bodies, while availability is limited to monthly values for others. There are on average 6000 datapoints per dataset, ranging from a minimum of 3000 to a maximum of 8000. A first glance reveals that on average 60% of the data contains missing values, indicating that severe preprocessing might be necessary.

4 Solution statement

The task at hand it to build a predictive model which predicts either water level or water flow for several water bodies. Since the type of features and target depend on the type of water body, a different model needs to be made for each type of water body. As the differences between waterbodies of the same type can be large, and the availability of features differs, it might be necessary to create a separate model for each individual water body.

As historic water level and water flow measurements are recorded it is possible to approach this task using supervised learning. In order to capture complex relations between features and target, a regression model which can capture non-linear relations might be preferred, such as Sklearn's RandomForestRegressor or MLPRegressor. Sklearn's common model interface and usage of pipelines allows us to easily explore several models. Furthermore, XGBoost, which has proven to be a winning algorithm in many previous Kaggle competitions, could offer a solution which is worth exploring.

The data at hand is not clean, as already mentioned in the *Datasets and inputs* section. The challenge is therefore expected to be mostly in the preprocessing and feature engineering, and less in the choice of the correct model.

5 Benchmark model

As a benchmark model an ordinary least square fit will be made using Sklearn's LinearRegression. This model with capture only linear relations between features and target.

6 Evaluation Method

Model evaluation will be done by R^2, in order to provide a measurement of the amount of variance captured by the features. In order to get a feel for the size of the errors, additionally the $RMSE$ and MAE will be used. A comparison between the values of $RMSE$ and MAE can tell something about the type of errors the model makes: many small errors or a few large errors.

In order to compare models of the same type applied to different datasets the $MAPE$ will be used. As this concerns a relative error, results on different datasets where target values can be different in orders of magnitude can be compared. This way it might be possible to find the type of model which works best for all water bodies (of one type), and have a more uniform approach across all water bodies to forecast water level and flow.

7 Project Design

The following steps outline the proposed workflow. Most of the work will be done in jupyter notebooks. Where possible, functions will be stored in a separate python file and loaded into the jupyter notebook. If time allows, the solution will be transferred to AWS Sagemaker for practise purposes. The size of the data does not require additional cloud resources and the work can be easily done on a laptop.

1. Data exploration: Familiarise oneself with the data
2. Data cleaning: Dealing with missing values, setting types of columns correctly, etc
3. Feature engineering: If necessary create more features based on the provided data, for instance by using lagged values. Encoding categorical features.
4. Feature selection: Select most promising features based on for example (ϕ_k) correlation, predictive power score
5. Create model evaluation code: to be used to compare different models
6. Create baseline solution: To be used as benchmark
7. Model exploration: Build a sklearn pipeline to easily explore several models
8. Create model training code: Model training and hyperparameter tuning
9. Create inference code

(10.) Optional: Implement Sagemaker model training and inference

References

[1] Acea Group
https://www.gruppo.acea.it/en

[2] Kaggle competition Acea Smart Water Analytics
https://www.kaggle.com/c/acea-water-prediction