Exit Report of Project for AquaFlow Technologies

MLOps course, Reichman University
February, 2024
Students names: Shahar Ehrenhalt, Oren Avida, Mayan Stroul, Alexander Gorelik

Customer: AquaFlow Technologies
Team members:

- Supervisor:
    - Ishai Rosenberg, Reichman University, MLOps Course Instructor
- Machine Learning and Data Science M.Sc. Students:
    - Shahar Ehrenhalt
    - Oren Avidan
    - Mayan Stroul
    - Alexander Gorelik
- Client - AquaFlow Technologies:
    - Data Administrator
    - Data Scientist (baseline owner)
    - CFO – to assess to provide inputs about financial costs
    - COO – operating the system and handling the alerts and regulation

Overview

This project aimed to refine AquaFlow Technologies' anomaly detection capabilities within their water treatment systems. Leveraging machine learning models LightGBM and MSET, and introducing an innovative SHAP-based feature selection methodology, the focus was on significantly reducing false negatives to mitigate operational and financial risks associated with undetected water quality anomalies.

Business Domain

The business domain is the Water Treatment Industry, specifically focusing on enhancing the reliability and efficiency of anomaly detection in water filtration systems.

AquaFlow Technologies (the “company”), a leader in the water treatment industry, is facing a challenge with predicting anomalies in water quality, using the data generated by its pumps and filtration units.

Business Problem

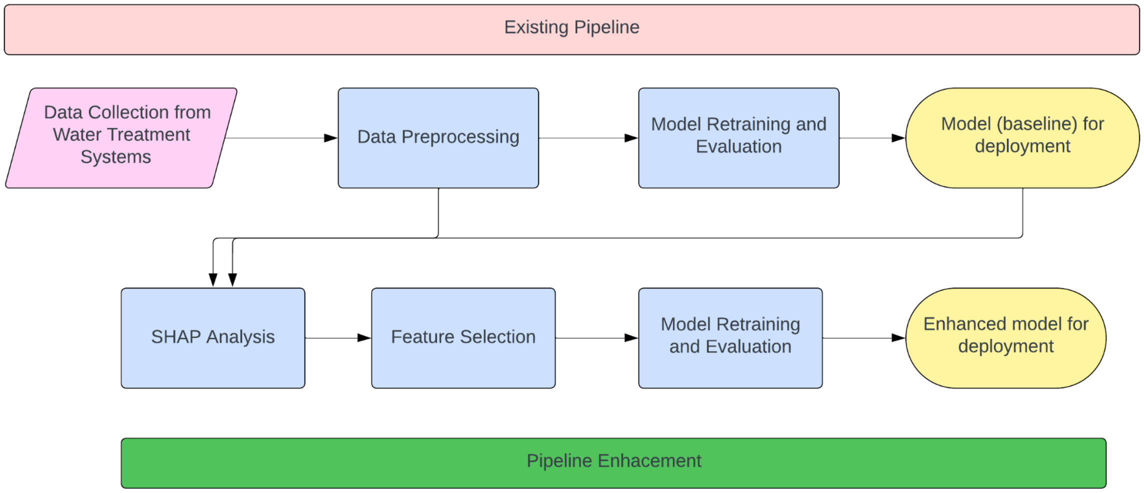
AquaFlow Technologies aimed to reduce false negatives in anomaly detection due to the increased financial and operational costs associated with undetected anomalies in water quality. The goal was to improve the system's recall without substantially increasing false positives (decreasing the precision score), balancing safety standards and operational efficiency against the backdrop of stringent regulatory compliance and heightened financial implications of missed anomalies.

Data Processing

The process began with 8-dimensional time series data from the water filtration systems' sensors. Data preprocessing included smoothing, standardization, and restructuring to create a 24-feature dataset optimized for machine learning analysis.

Solution Architecture

The solution architecture integrated data collection, preprocessing, model training, SHAP analysis for feature optimization, model retraining, and re-evaluation.
This architecture was chosen in order to enhance F-beta score and improve the performance of the pipeline.


Existing Pipeline
Data Collection from Water Treatment Systems
Data Preprocessing
Model Retraining and Evaluation
Model (baseline) for deployment
SHAP Analysis
Feature Selection
Model Retraining and Evaluation
Enhanced model for deployment
Pipeline Enhacement

Benefits

Company Benefit

The project enriched the team's expertise in machine learning model development and feature selection techniques, especially with time series data, contributing to the internal knowledge base and capabilities in anomaly detection solutions.

Customer Benefit

For AquaFlow Technologies, the refined anomaly detection system promises increased operational reliability and efficiency, potentially leading to cost savings in maintenance and reduced downtime. Our model was able to decrease 9 missing anomalies during the test period which lasts 3 hours, 44 minutes and 40 seconds. This period represents the real data distributiom. The potential of missing one anomaly is up to 10000$, which means our solution can reduce about $210,240,000 a year.

Learnings

Data science / Engineering

The project highlighted the effectiveness of SHAP analysis in improving model accuracy by optimizing feature selection, particularly in the context of anomaly detection.
It also improved our knowledge in how to work with time series data.

Domain

Gained insights into the specific challenges of anomaly detection in water filtration systems, including the variability of sensor data and the critical importance of minimizing false negatives.

Product

Leveraged LightGBM's and MSET’s capabilities for efficient model training on large datasets, and explored the application of SHAP for feature analysis.

What's unique about this project, specific challenges

Since the classic SHAP solution has not produced improved results to the baseline model, we had to think about a different solution. We then came up with the idea of heuristic SHAP, which was able to improve the results and reduce false negatives.

Links

Project repository, Github:
https://github.com/goreliks/MLOps_RUNI_MSc/tree/main

YOSHI_MANAMINORISA- Anomaly Detection using LightGBM, Kaggle:
https://www.kaggle.com/code/yoshimanaminorisa/anomaly-detection-using-lightgbm

waico/SKAB, Github:
https://github.com/waico/SKAB/blob/master/notebooks/MSET.ipynb

Next Steps

The company will continue the exploration of additional data sources and machine learning techniques such as the following:

- Magnitude use- Collect the magnitude of each feature into account and not only the direction of each feature
- Optimization- Adding the optimization step for optimal number of features selection for drop
- OmniXAI- Using "temporal context" - explaining a sample by features from other timestamp as well
- Other ML techniques.

The mentioned above will continue to improve the anomaly detection capabilities. Regular follow-up meetings are planned to monitor the system's performance and ensure its alignment with operational needs and regulatory changes.

Appendix

Baseline model report:
https://github.com/goreliks/MLOps_RUNI_MSc/tree/main/reports/baseline.pdf

Final Model Report:
https://github.com/goreliks/MLOps_RUNI_MSc/tree/main/reports/final_model_report.pdf