Baseline Model Report- AquaFlow Technologies

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

The baseline models for this project are the LightGBM (Light Gradient Boosting Machine, a gradient-boosting framework that uses tree-based learning algorithms) and the MSET (Multivariate State Estimation Technique). Both models are designed to detect anomalies in time series data generated from the measuring units of AquaFlow Technologies. In this report we will focus on the LGBM model as required.

The raw data, which is 8-dimensional and obtained from the sensors, undergoes several preprocessing steps including visualization, splitting, smoothing, and standardization. The model is then trained on this preprocessed data.

This baseline model serves as a starting point for our anomaly detection improvement task. It provides a benchmark against which we can compare the performance of our improved model. The ultimate goal is to enhance the model's ability to detect anomalies accurately, specifically minimizing false negatives, thereby contributing to the efficiency and reliability of the filtering water system.

Analytic Approach

- **Target Definition:** The target is to decrease False-Negative predictions in the company's current anomaly detection system, using time series data of a water filtration system offered by AquaFlow Technologies. The anomalies are labeled in the data (normal: 0, anomaly: 1).
- **Inputs:** The raw data is an 8-dimensional time series data obtained from the company's water filtration systems. The features include Accelerometer1RMS, Accelerometer2RMS, Current, Pressure, Temperature, Thermocouple, Voltage, and Volume Flow RateRMS (After preprocessing, the input data includes 24 features).
- **Baseline Model:** LightGBM model was built for anomaly detection. The model was trained using the training data.

Model Description

As a baseline model, we used the LightGBM model as given. LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework developed by Microsoft that uses tree-based learning algorithms. It's designed for speed and efficiency with the capability of handling large datasets. LightGBM is popular for its performance and lower memory usage compared to other gradient boosting frameworks.

The data flow involves several steps including data loading, visualization, preprocessing (splitting, smoothing, standardization), and finally, model building and anomaly detection.

The notebook performs data preprocessing for time-series analysis by first smoothing the data using a Kaiser window method to reduce noise. It then standardizes the smoothed data to have a mean of zero and standard deviation of one for all features. Next, it restructures the data into a format suitable for analysis by creating new datasets that summarize the past observations with mean, minimum, and maximum values within a specified look-back window.

The hyperparameters for this model include learning rate, min_data_in_leaf, max_depth, and num_leaves. The baseline model includes optimization of those hyper-params.

Results (Model Performance)

Summary of the model results as seen in the Jupyter Notebook:

```
[LightGBM] [Info] Number of positive: 4309, number of negative: 8394
[LightGBM] [Info] Total Bins 6120
[LightGBM] [Info] Number of data points in the train set: 12703, number of used
features: 24
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.339211 -> initscore=-0.666811
[LightGBM] [Info] Start training from score -0.666811
***************************************
TEST ACCURACY: 0.9214
TEST F1 SCORE: 0.9008
TEST FBETA SCORE: 0.9578
TEST CONFUSION MATRIX:
array([[1041,    0],
       [ 145,  658]])
```

Summary- Confusion Matrix:

LGBM (baseline model)	Predicted Non-Anomaly	Predicted Anomaly
True Non-Anomaly	**1041**	**0**
True Anomaly	**145**	**658**

As we can see, and following the mentioned in the mid-term project, the baseline model achieved F1 score and F Beta score of 0.9008 and 0.9578 respectively, which indicates good results. As mentioned, the company already has a relatively good system for anomaly detection.

Model Understanding

The model was able to distinguish between normal and anomaly classes in the dataset effectively. The model's high F1 score suggests that it has a good balance of precision and recall, making it a reliable tool for anomaly detection. As we can see, the model has 0 FPs, but the model still needs some improvements in the FNs detection process.

Our hypothesis was whether all features used in the prediction process contributed to its performance, and whether removing some of the features (or having a better feature selection process) could improve the results.

Conclusion and Discussions for Next Steps

The project successfully implemented a LightGBM model for anomaly detection in time series data from a water filtration system. The model achieved an F1 score of 90% and F beta score of 96%, indicating good balance in terms of recall and precision.

Although the results indicate good results we want to improve them because of our business reasons. In the improved model, we will use SHAP to assess the contribution of each of the features, and we expect to see an improvement in the model performance after a better feature selection process.

Other relevant data sources could potentially include other datasets of similar machinery or systems, provided they also include sensor readings and anomaly labels.