MODEL CARD

The model has been developed to predict the level flow data points collected by Level sensors attached to Remote Telemetry Loggers. Occasionally, these sensors record inaccurate readings, such as negative values or erroneous spikes, caused by reflections during the measurement process. Although the current data accuracy of the sensors with this model reaches up to 94%, and in some devices, even 98%, emerging analytical software that requires 100% accuracy faces processing issues when data quality falls below this threshold. Consequently, a model is needed to predict and correct occasional negative values and spikes by using the sensor's historical and forward data, as well as historical and forecasted rainfall data, to fill in the missing or erroneous data. This is where Bayesian Optimization is integrated into the model to minimize errors and improve the quality of the sensor data.

Person developing the model: Palzor Lama
Model date: 23/09/2024
Model version: Predict Missing Values v1.01
Model type: Bayesian Optimization

Intended Use:

- **Primary Intended Uses:** The model ensures that the level flow data collected from sensors is more accurate and reliable, making it suitable for monitoring, analyzing, and predicting spillages, overflows, and blockages in wastewater systems. This ensures robust and resilient analysis for future event predictions.
- **Primary Intended Users:** The intended users are Water Utilities operating in the Wastewater Sector, particularly at Combined Sewer Overflow Sites, as well as sites in the Clean Water Sector.
- **Out of Scope Use Cases:** The model can also be adapted for pressure sensors that exhibit similar noise and erroneous spikes due to system conditions. Additionally, it can be used for other sensor setups such as soil moisture probes and weather stations.

Metrics: The model uses Mean Squared Error (MSE) during the Bayesian Optimization process to evaluate the difference between predicted and actual values.

Inputs: The model evaluates the data based on Level Data from Remote Telemetry Loggers, which trend every 15 minutes under normal conditions and every minute during alarm events. The model also incorporates rainfall data, including historical, real-time, and forecast data, to improve predictions.

Outputs: The output of the model is the corrected level flow data, with all negative values and erroneous spikes removed.

Evaluation Data: The evaluation involves plotting the corrected data against the original data, demonstrating the improvements achieved by applying the model.

Training Data: The training data consists of previous and forward data points from the Remote Telemetry Logger, combined with historical, real-time, and forecasted rainfall data to improve prediction accuracy.

Testing Data: The model currently uses the same previous and forward data points for testing. In future testing phases, side-by-side comparisons with additional Remote Telemetry Loggers could be performed to validate the model's accuracy further.

Quantitative Analysis:

The model aims to enhance data accuracy from an initial 94% to nearly 100%. It integrates rainfall data with the logger's flow measurements, refining the model's predictions and performance through continuous Bayesian Optimization.

Ethical Considerations:

- **Bias:** The model relies heavily on data from the Remote Telemetry Logger and rainfall forecasts. If either source contains errors or inaccuracies, the model may inherit these issues, potentially affecting its predictions. There is currently no step to validate if logger data or forecasts are faulty.
- **Critical Data:** The model must not suppress critical data, such as a spillage or blockage that results from extreme weather or other conditions. As a precaution, the model is tuned not to predict data if there are more than four consecutive spikes, ensuring the model doesn't overlook actual event data that could be crucial.

Caveats and Recommendations:

- **Model Evaluation:** The model has been tested using real site data and MET Office rainfall datasets. It successfully identifies negative values and erroneous spikes. The model's built-in tuning mechanism, which refrains from predicting values if more than four spikes are detected, ensures it captures actual spikes while avoiding unnecessary corrections.
- **Future Improvements:** The model can be further enhanced by incorporating additional rainfall datasets or more sophisticated rainfall models to improve prediction accuracy. Further in-house testing using controlled environments or more extensive telemetry logger data can help fine-tune the model. This would allow for even better accuracy and reliability.
- **Deployment Recommendations:** Currently, the model is designed to run on a cloud server. However, in the future, it could be deployed directly on the logger itself, enabling real-time predictions at the edge, further improving responsiveness and accuracy in dynamic environments.

Conclusion:

This model significantly enhances the accuracy of level flow data collected from Remote Telemetry Loggers, addressing the challenge of missing or erroneous data points caused by sensor reflections. With the application of Bayesian Optimization, the model has successfully increased

data quality from 94% to nearly 99.9%. This improvement will greatly benefit analytical software that relies on highly accurate data, especially in critical systems like wastewater monitoring. The model also demonstrates versatility, with potential applications across various sensors, further advancing the Remote Telemetry Loggers toward edge computing solutions. By integrating rainfall data and avoiding over-correction for spike events, the model ensures reliability and robustness in real-time environments, offering a scalable solution for predictive analytics in the water industry.