OpsFree Data Science Intern Assignment

Submitted by :

Sabhavath Yashwanth

Yashwanthsabhavath98.8@gmail.com

9441681145

#2. Algorithms to find anomalies in the duration of a span

Identifying outliers in the dataset is very crucial step in data analysis for better model building otherwise overfitting, imbalanced data, underfitting problems may arise which restricts the performance of a model prediction. Model can't be trained properly and results inaccurate prediction.

Algorithms to find Anomalies

Statistical methods include - **Z-Score and Inter-Quartile Range** methods for numerical dataset**
Machine Learning approaches include - **Isolation Forest** in which Splits are chosen randomly, aiming to isolate points quickly, **SVM** algorithm, **Ensemble methods** and **Autoencoders** : Anomalies are detected based on reconstruction error, where higher errors indicate anomalies.
Time Series approaches include : **LSTM, ARIMA AND SARIMA**
Additionally **BERT and Transformers** can also be used in detecting anomalies in microservice tracing data.

- Time Series BERT (T-BERT): BERT can be adapted for time series by modifying its input representation to handle continuous numerical data instead of discrete tokens.
- Encoder-Decoder Architecture and Self-Attention Mechanism: This can be used to model the time series data where the encoder captures the temporal dependencies and the decoder predicts future values. Anomalies can be detected based on the deviation between predicted and actual values.
- **Conclusion:**

Inter-Quartile Range or Z-Score would be a better choice for identifying the Outlier's and **Box-plot** for visualization.

For Detecting Anomalies in the microservice data we can implement following algorithms :

1. Time-Series Anomaly Detection : **LSTM**, **ARIMA**, **SARIMA.**
2. Machine Learning approaches : **Isolation Forest, SVM, Ensemble learning.**
3. Deep Learning Approaches : **Autoencoders, Transformers and BERT.**