Anomaly Detection for Electric Vehicles – Aaryan (102353014)

Exploratory Data Analysis

The dataset contained two main components:

- Normal operating data (100,001 records, 9 features)
- Attack data with anomaly labels (90,005 records, 10 features)

Key EDA Insights

1. **Data Distribution**:
   - Normal data showed consistent patterns with stable RPM (~2500) and speed (~29 units)
   - Attack data showed bimodal distributions for key parameters like RPM and speed, with two distinct operating modes (normal range and anomalous range ~6800 RPM/80 units speed)
2. **Anomaly Labeling**:
   - 37.81% of the attack dataset was labeled as anomalies (34,027 samples)
   - The anomalies showed significantly different statistical patterns compared to normal data
3. **Feature Correlations**:
   - Strong correlations between RPM and speed (r=1.00)
   - Strong correlations between Actuator_Pos_1 and Throttle_com (r=0.98)
   - In attack data, anomaly labels strongly correlated with RPM, speed, and Throttle_com (r=0.99)
4. **Missing Values**:
   - Both datasets contained minimal missing values (5 in normal data, 6 in attack data)
   - No duplicate records were found
5. **Feature Differences During Anomalies**:
   - RPM: Normal mean=2533.68, Anomaly mean=6769.77

    - Speed: Normal mean=29.85, Anomaly mean=79.75
    - Throttle_com: Normal mean=0.05, Anomaly mean=0.54

Models Implemented

1. Isolation Forest

Isolation Forest is an unsupervised learning algorithm that detects anomalies by isolating observations through random feature selection and random split values.

Implementation Details:

- Contamination parameter set to match class distribution (0.4737)
- 100 estimators used for robustness
- Features standardized before model training

2. LSTM Autoencoder

The LSTM Autoencoder learns to reconstruct normal behavior patterns and flags data points with high reconstruction error as anomalies.

Implementation Details:

- Sequence length: 5 time steps
- Architecture:
    - Encoder: LSTM(64) → Dropout(0.2) → LSTM(32) → Dropout(0.2)
    - Decoder: Dense(32) → Dense(40) → Reshape to original dimensions
- Trained on normal data only (80,000 samples)
- Anomaly threshold set at 95th percentile of reconstruction error on normal data (1.361749)

Model Evaluation and Comparison

Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Isolation Forest	0.9031	0.8957	0.9003	0.8980
LSTM Autoencoder	0.5533	0.6697	0.1127	0.1929

Confusion Matrix Analysis

Isolation Forest:

- True Negatives: 18,113 (correctly identified normal)
- False Positives: 1,888 (false alarms)
- False Negatives: 1,795 (missed anomalies)
- True Positives: 16,206 (correctly identified anomalies)
- False Alarm Rate: 0.0944
- Miss Rate: 0.0997

LSTM Autoencoder:

- True Negatives: 18,998 (correctly identified normal)
- False Positives: 1,000 (false alarms)
- False Negatives: 15,972 (missed anomalies)
- True Positives: 2,028 (correctly identified anomalies)
- False Alarm Rate: 0.0500
- Miss Rate: 0.8873

Conclusion

The Isolation Forest model significantly outperformed the LSTM Autoencoder for this EV anomaly detection task. While the LSTM Autoencoder had fewer false positives (better precision), it missed a large number of actual anomalies (poor recall), making it less suitable for security applications where detecting all potential attacks is critical.

The strong performance of Isolation Forest suggests that the anomalies in this EV dataset are well-separated from normal behavior in the feature space, making them easier to detect with distance-based methods than with reconstruction-based approaches. For real-world implementation, the Isolation Forest model would be recommended due to its higher F1 score, better recall, and simpler implementation.

Future work could explore ensemble methods combining both approaches or feature engineering to improve the LSTM model’s performance on the temporal aspects of the data.