DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

Final Exam Report - Q1 Supervised Outlier Detection

50015940 Jiaxiang Gao

Main Experimental Steps and Methods:

1. Data Preparation:

Feature and Target Separation: The feature variables (X) were separated from the target variable (Is_Falling) in the training dataset.

Feature Standardization: The StandardScaler was utilized to scale the features, ensuring a uniform evaluation.

Handling Class Imbalance: SMOTE (Synthetic Minority Over-sampling Technique) was applied to generate synthetic samples for the minority class, thus balancing the dataset.

2. Test Data Preparation:

Test Data Processing: The test dataset was processed by dropping the 'ID' and 'Is_Falling' columns. The 'Is_Falling' column is the target variable to be predicted, and 'ID' is not a feature. The same scaler used for the training data was applied to scale the test data.

3. Model Training and Evaluation:

3.1 Random Forest Classifier:

Training: A RandomForestClassifier was trained using the resampled training data.

Evaluation: The classifier's performance was evaluated on the test data, focusing on metrics like precision, recall, and F1-score.

3.2 LightGBM Classifier:

Training: An LGBMClassifier was similarly trained on the resampled training data.

Evaluation: This classifier's performance was also assessed using the same metrics as the RandomForestClassifier.

4. Comparison and Decision:

Performance Analysis: The recall rate for the minority class (falling class) was a key metric in the comparison. LightGBM demonstrated a significantly higher recall rate compared to the Random Forest, indicating its superior ability in identifying true falling events.

Random Forest Classifier Evaluation

	precision	recall	f1-score	support
0	0.96	0.96	0.96	6280
1	0.25	0.26	0.25	343
accuracy			0.92	6623
macro avg	0.61	0.61	0.61	6623
weighted avg	0.92	0.92	0.92	6623

LightGBM Classifier Evaluation

	precision	recall	f1-score	support
0	0.97	0.91	0.94	6280
1	0.25	0.52	0.33	343
accuracy			0.89	6623
macro avg	0.61	0.72	0.64	6623
weighted avg	0.93	0.89	0.91	6623

Final Selection: Based on the higher recall rate for the minority class, LightGBM was chosen for predicting the test data, especially considering the critical nature of the application (fall detection).