Q1. Supervised Outlier Detection

1. experimental steps

First, we import the dependencies. In this task we use pandas, sklearn, imblearn and xgboost.

Then we read the file. In this part we will use os to get the path of all CSV files in the folder and the merge them into a dataframe.

Next, we use two different methods to solve the problem. We take a part of training set to be a validation set. After training and validating the models, we evaluate them on test set. The better result would be the ouput.

Finally, we write the output as csv and put it into folder.

2. Introduction to the methods

Method 1: oversampling

Oversampling is a technique used in machine learning to address class imbalance problems. Class imbalance occurs when the number of instances in one class is significantly higher or lower than the number of instances in another class. This can lead to biased models that are more accurate in predicting the majority class but perform poorly on the minority class.

One popular oversampling method is Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by creating synthetic instances of the minority class by interpolating between neighboring instances. It randomly selects an instance from the minority class and finds its k nearest neighbors. It then generates new instances by randomly selecting one of the neighbors and creating a synthetic instance along the line segment joining the original instance and the selected neighbor.

SMOTE-Tomek combines the SMOTE oversampling technique with the Tomek links undersampling technique. Tomek links are pairs of instances from different classes that are closest to each other, and their removal can help improve the separation between classes.

SMOTE-Tomek first applies SMOTE to oversample the minority class, and then uses Tomek links to

remove instances that are identified as being close to instances from the majority class. This combined method helps to further balance the classes by removing instances that could potentially cause misclassification or introduce noise.

Method 2: Ensemble model

The Ensemble method utilizes techniques such as multi model combination, voting mechanism, weighting mechanism, and anomaly detection to balance the number of samples from different categories in imbalanced datasets, increase attention to minority classes, and thus improve the predictive accuracy of the model.

XGBoost (eXtreme Gradient Boosting) is a popular and powerful machine learning algorithm that is widely used for both regression and classification tasks. It is an implementation of the gradient boosting framework, which combines multiple weak prediction models (typically decision trees) to create a strong predictive model.

XGBoost has some features and techniques for imbalanced datasets, which can help improve the accuracy of models in predicting minority classes. Overall, XGBoost has flexibility and adjustability in handling imbalanced datasets, and can improve the model's predictive ability in minority classes through techniques such as weight adjustment, learning rate reduction, sampling methods, classification threshold adjustment, and objective function adjustment. This makes XGBoost an effective choice for handling imbalanced datasets.