COURSE NAME: Hillel - Machine Learning. Lesson 16 (06.04.2023 19:15,) Outlier Detection
LECTURER: Yevhen Krasnokutsky
STUDENT: Maryna Lysenko

HW 13 REPORT “ Outlier Detection”.

Based on dataset of your choice test metric (accuracy / MSE) for training:

- Without filtering-out outliers
- With filtering-out outliers

Датасет: Mushroom Classification - классификация съедобен ли гриб
https://www.kaggle.com/datasets/uciml/mushroom-classification

Доля несьедобных семплов в датасете 48,1%, данные достаточно сбалансированы.
Shape of mushrooms data (8124, 23)

Влияние удаления выбросов

Натренировано несколько моделей, с которыми удалось достичь точности на уровне 100%: CatBoostClassifier, LogisticRegression, DecisionTreeClassifier, SVC

Модель RandomForestClassifier - 0.99836

Модель XGBClassifier - 0.99836

Последнюю выбрано для оценки влияния удаления шумов из данных

IsolationForest:

С удалением 10% - дало точность хуже 0.99549

На также очищенной от выбросов тестовой выборке (это дает смещение в распределении данных, некорректно будет сравнение с базовой моделью) точность выросла до 0.99864

С удалением 0.5% - тоже точность хуже и составляет 0.99549

OneClassSVM

nu=0.01

Accuracy Score is 0.99713

Точность хуже базовой, но лучше чем по предыдущему методу результаты

Conclusions:

1. В случае с этой задачей несъедобный гриб скорее всего и должен иметь какие-то назовем это “дефекты и шумы” и это должно бы больше свидетельствовать о принадлежности к целевому классу. Возможно поэтому не удалось достичь прироста при использовании очищения выборки от выбросов.

2. В целом эксперимент показывает, что удаление выбросов не всегда может быть необходимым, особенно если набор данных хорошо сбалансирован, а используемые модели устойчивы к выбросам. Однако, если выбросы присутствуют и их влияние на производительность модели значительно, методы обнаружения и удаления выбросов могут быть полезны для повышения точности модели.