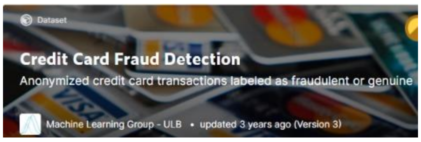
OASIS ML Group TRAINING 09



Dataset

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

Machine Learning Group - ULB • updated 3 years ago (Version 3)

◆ Let’s deal with Imbalanced datasets

Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce errors. However, if the data set in imbalance, then in such cases, you get a pretty high accuracy just by predicting the majority class, but you fail to capture the minority class, which is most often the point of creating the model in the first place. **“Credit card fraud detection”** is a famous example of imbalanced dataset. The dataset is download from Kaggle and it contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where it has **492 frauds out of 284,807** transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Please try to build an efficient model which can find out whether the credit card transaction was fraudulent or not.

▫ Practice：

- **Data Analysis**
- **How to handle imbalanced datasets**

  There’re lots of issues to discuss in this topic, so please refer to the website provided below and **finish the task**. **Share what you find as more as you can** and try your best to convince your model can make the prediction more precise.

▫ Reference：

- Kaggle - Credit Card Fraud Detection
- 知呼 - Imbalance data 分析 - Credit Card Fraud Detection
- 知乎 - 閾值/特徵篩選分析 - Credit Card Fraud Detection
- CSDN - Credit Card Fraud Detection Spark 建模
- CSDN - 解决不平衡数据集问题
- CSDN - 一文教你如何處理不平衡数据集
- Imbalanced Classification with the Fraudulent Credit Card Transactions Dataset
- Toward data science - Credit Card Fraud Detection
- Toward data science - Credit Card Fraud Detection Using Machine Learning & Python
- Toward data science - Handling imbalanced datasets in machine learning
- Section - Handling Imbalanced Datasets in Machine Learning