Data Sheet for Credit Card Fraud Detection Dataset

# Motivation

Purpose: This dataset was created to detect fraudulent credit card transactions using machine learning models. The goal is to address the imbalance between fraudulent and non-fraudulent transactions and improve fraud detection accuracy.  
Dataset creators: The dataset was provided by the Machine Learning Group of ULB (Université Libre de Bruxelles).  
Funding: Not specified in the source.

# Composition

Instances: The dataset contains anonymized credit card transaction records, with each instance representing a transaction.  
Total instances: 284,807 records.  
Labels: The dataset is labeled with two classes: '0' (non-fraudulent) and '1' (fraudulent).  
Features: 30 features including 'Time', 'Amount', and 28 anonymized features (V1 to V28) derived via PCA.  
Confidentiality: The dataset does not contain personally identifiable information (PII).  
Data split: The dataset was split into training (80%) and testing (20%) sets. The split preserves the class imbalance.

# Collection Process

Acquisition: The data represents real-world credit card transactions collected over a period of time, but the specific acquisition details are not provided.  
Sampling: The dataset is imbalanced, with 492 fraud cases and 284,315 non-fraud cases.  
Time frame: Transactions collected over two days in September 2013.  
Ethics: Anonymized data ensures privacy and confidentiality.

# Preprocessing/Cleaning

Preprocessing: Features were scaled, and the dataset was resampled using SMOTE (Synthetic Minority Over-sampling Technique) to balance the class distribution.  
Raw data: The raw data includes both non-fraudulent and fraudulent transactions, and features are anonymized.

# Uses

Potential uses: Fraud detection, anomaly detection, research on imbalanced datasets.  
Risks: Since the dataset is anonymized, there are no privacy concerns, but care must be taken when applying the model in other domains to avoid bias.  
Restrictions: The dataset should not be used to derive any sensitive information about individuals.

# Distribution

External distribution: The dataset is publicly available on Kaggle, under a public license.  
License: The dataset is distributed under the CC0: Public Domain license.

# Maintenance

Responsibility: The dataset is maintained by the Machine Learning Group of ULB.