Datasheet for Credit Card Fraud Detection Dataset

**Motivation**

* For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset has been collected and analysed during a research collaboration of Wordline and the Machine Learning Group (MLG) of ULB (Universite Libre de Bruxelles) (link: [http://mlg.ulb.ac.be](http://mlg.ulb.ac.be/)). This dataset was created to assist with a number of different research projects around big data mining and credit card fraud detection, that were conducted by the MLG ULB group. A list of all the projects that have used this dataset can be found in the Appendix section at the end of this document.

* Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset?

Wordline, which is a French multinational payment and transactional services company, created the dataset in collaboration with the MLG-ULB, as part of research project that aimed on improving the existing fraud detection processes of Worldline (<https://mlg.ulb.ac.be/wordpress/portfolio_page/defeatfraud-assessment-and-validation-of-deep-feature-engineering-and-learning-solutions-for-fraud-detection/>). Innovoris, the Brussels Region Institute for Research and Innovation, provided the funding through the Team Up Programme “*DefeatFraud: Assessment and validation of deep feature engineering and learning solutions for fraud detection”*, that took place from 2018 -2021.

**Composition**

* What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Please provide details.

The dataset consists of credit card transaction that were made in September 2013 by European cardholders. All transactions occurred in two days. The dataset contains contains only numerical input variables. Features V1-V28 are the outputs of a principal component analysis (PCA )transformation. No information is provided about the original features due to confidentiality issues. Feature ‘Time’ contains the seconds elapsed between each transaction and the first transaction in the dataset. Feature ‘ Amount’ is the transaction amount, but no information on the currency are given. Finally, the dataset includes a ‘Class’ feature, which is the response variable and takes the value 1 if a transaction is fraudulent otherwise 0.

* How many instances of each type are in total?

There are 284807 instances.

* Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

These information are unknown to the author of the datasheet.

* What does each instance consist of? Raw data? Unprocessed? Text, images?

Numerical data for each one of the 31 features, described above.

* Are there any labels to the data?

Yes, the labels are binary. Label 1 represents a fraudulent transaction while 0 a normal transaction.

* Is there any missing information from individual instances?

No missing instances in the dataset.

* Are relationships between individual instances made explicit?

None explicitly.

* Are there recommended data splits (e.g. train / test)? Provide a description of the splits, and the rationale behind them.

No data split is provided by the source of the dataset. However, it is provided the information that the dataset is highly imbalanced with only 0.172% of the entries belonging to the positive class (‘fraud’) and this should be taken into consideration when performing data splits.

* Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

* Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.

The dataset contains information about card transactions, which is usually confidential. However, all the provided data are an output of a principal component analysis (PCA), in order to protect confidentiality.

* Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

* Does the dataset identify any subpopulations (e.g., by age, gender)?

No.

* Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

* Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Yes, It mostly consists of financial data.

**Collection process**

* How was the data associated with each instance acquired? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

These information are unknown to the author of the datasheet.

* If the data is a sample of a larger subset, what was the sampling strategy? Deterministic, random, etc...?

These information are unknown to the author of the datasheet.

* Over what time frame was the data collected?

It is stated that the data consists of transactions that happened in 2 days in September 2013, but no more information are given.

* Were there any ethical review processes conducted (e.g. by an institutional reviewing board?)

These information are unknown to the author of the datasheet.

* Were the individuals notified of the collection of the data?

These information are unknown to the author of the datasheet.

* Did the individuals consent to their data being collected?

These information are unknown to the author of the datasheet.

* If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

These information are unknown to the author of the datasheet.

* Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

These information are unknown to the author of the datasheet.

**Preprocessing/cleaning/labelling**

* Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

As mentioned in previous section the data from this datasheet are the output of a PCA analysis, but we don’t have more information.

* Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

These information are unknown to the author of the datasheet.

**Uses**

* What other tasks could the dataset be used for?

These information are unknown to the author of the datasheet.

* Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Unknown to the authors of the datasheet, however, we cannot think of any obvious impacts.

* Are there tasks for which the dataset should not be used? If so, please provide a description.

Unknown to the authors of the datasheet.

**Distribution**

* Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available.

* How will the dataset be distributed?

It is distributed by Kaggle (link: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>). There might also be other places where the dataset is available.

* When will the dataset be distributed?

It is currently available.

* Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is distributed under the Database Contents License (DbCL) v1.0 license (<https://opendatacommons.org/licenses/dbcl/1-0/>)

**Maintenance**

* Who will be maintaining the dataset?

Unknown to the authors of the datasheet.

**Appendix**

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. [Calibrating Probability with Undersampling for Unbalanced Classification.](https://www.researchgate.net/publication/283349138_Calibrating_Probability_with_Undersampling_for_Unbalanced_Classification) In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. [Learned lessons in credit card fraud detection from a practitioner perspective](https://www.researchgate.net/publication/260837261_Learned_lessons_in_credit_card_fraud_detection_from_a_practitioner_perspective), Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. [Credit card fraud detection: a realistic modeling and a novel learning strategy,](https://www.researchgate.net/publication/319867396_Credit_Card_Fraud_Detection_A_Realistic_Modeling_and_a_Novel_Learning_Strategy) IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea [Adaptive Machine learning for credit card fraud detection](http://di.ulb.ac.be/map/adalpozz/pdf/Dalpozzolo2015PhD.pdf) ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. [Scarff: a scalable framework for streaming credit card fraud detection with Spark](https://www.researchgate.net/publication/319616537_SCARFF_a_Scalable_Framework_for_Streaming_Credit_Card_Fraud_Detection_with_Spark" \t "_blank), Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. [Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization,](https://www.researchgate.net/publication/332180999_Deep-Learning_Domain_Adaptation_Techniques_for_Credit_Cards_Fraud_Detection) International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi [Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection](https://www.researchgate.net/publication/332180999_Deep-Learning_Domain_Adaptation_Techniques_for_Credit_Cards_Fraud_Detection), INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi [Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection](https://www.researchgate.net/publication/333143698_Combining_Unsupervised_and_Supervised_Learning_in_Credit_Card_Fraud_Detection)Information Sciences, 2019

Yann-Aël Le Borgne, Gianluca Bontempi [Reproducible machine Learning for Credit Card Fraud Detection - Practical Handbook](https://www.researchgate.net/publication/351283764_Machine_Learning_for_Credit_Card_Fraud_Detection_-_Practical_Handbook)

Bertrand Lebichot, Gianmarco Paldino, Wissam Siblini, Liyun He, Frederic Oblé, Gianluca Bontempi [Incremental learning strategies for credit cards fraud detection](https://www.researchgate.net/publication/352275169_Incremental_learning_strategies_for_credit_cards_fraud_detection), IInternational Journal of Data Science and Analytics