About Dataset

Context

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Content

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Update (03/05/2021)

A simulator for transaction data has been released as part of the practical handbook on Machine Learning for Credit Card Fraud Detection - <https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_3_GettingStarted/SimulatedDataset.html>. We invite all practitioners interested in fraud detection datasets to also check out this data simulator, and the methodologies for credit card fraud detection presented in the book.

Acknowledgements

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group ([http://mlg.ulb.ac.be](http://mlg.ulb.ac.be/)) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

More details on current and past projects on related topics are available on <https://www.researchgate.net/project/Fraud-detection-5> and the page of the DefeatFraud project

Please cite the following works:

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. 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, 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, 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 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, 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, 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, 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 Information Sciences, 2019

Yann-Aël Le Borgne, Gianluca Bontempi Reproducible 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, IInternational Journal of Data Science and Analytics  
  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
Usefulness of the Project

This project demonstrates a comprehensive workflow for detecting fraud in credit card transactions, which has real-world applications in the financial industry. Here’s why it’s useful:

Fraud Prevention:

Detects fraudulent transactions with high accuracy, protecting financial institutions and customers.

Reduces financial losses and improves trust in the payment systems.

Scalability:

Techniques used (e.g., TabNet, advanced preprocessing) are scalable to large datasets, making them suitable for production environments.

Explainability:

By leveraging explainable models like TabNet, stakeholders can understand how features contribute to fraud detection, increasing the adoption of AI solutions in regulated industries.

Imbalanced Data Handling:

The project addresses class imbalance, which is a critical challenge in fraud detection, ensuring the model is not biased towards the majority class.

Advanced Evaluation Metrics:

Uses precision-recall AUC, MCC, and F1-score, which are more meaningful for imbalanced datasets than simple accuracy metrics.

Why It Is Advanced

This project is advanced due to the following reasons:

End-to-End Pipeline:

Covers the entire data science workflow: from EDA and feature engineering to advanced modeling and evaluation.

Sophisticated Data Preprocessing:

Implements outlier detection (Isolation Forest) and transformations (logarithmic, time grouping) to clean and enhance data.

Handles imbalanced data with SMOTE and ensemble techniques, ensuring robust predictions.

Feature Engineering:

Generates interaction and derived features, extracting more value from the dataset for better predictive performance.

Cutting-Edge AI Models:

Utilizes TabNet, a state-of-the-art deep learning model specifically designed for tabular data, combining feature selection and learning in one step.

Explainable AI (XAI):

TabNet inherently provides interpretability, making it suitable for applications where understanding the decision-making process is crucial.

Efficient Training:

Implements early stopping with epoch thresholds, ensuring the model trains only as much as needed, saving resources and preventing overfitting.

Comprehensive Evaluation:

Uses multiple performance metrics, ensuring the model’s reliability across different scenarios and imbalanced datasets.