earch kaggle

Competitions

Datasets

Kernels

Discussion

Jobs



Credit Card Fraud Detection



Anonymized credit card transactions labeled as fraudulent or genu...



by Andrea · last updated 2 months ago

Overview

Kernels

Discussion

5 days ago

Activity

Download (68 MB)

New Notebook

Simple LR Accurac...
run 18 days ago

Starter Kernel (abo...
run 2 months ago

In depth skewed da...
run 7 days ago

Votes

Discussion

In depth skewed d...
a day ago

Detects 83% of fra...
4 days ago

Simple LR Accurac...

2

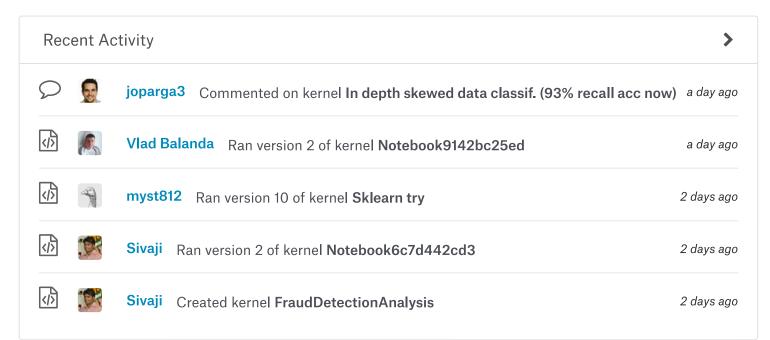
replies

Top Contributors

BrunoGuilh... 1st

Nandy Sin... 2nd

joparga3 3rd



The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset present 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.

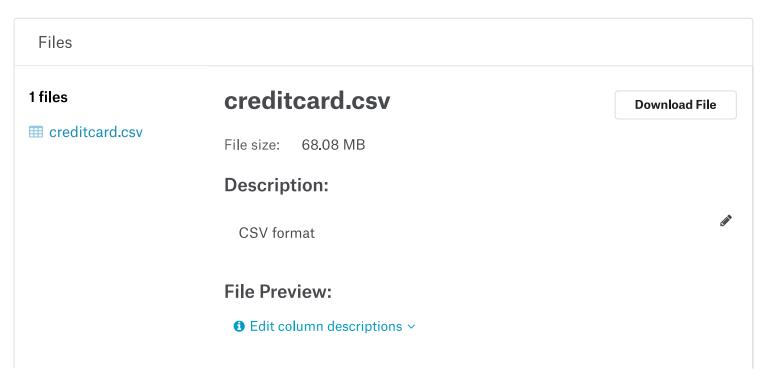
It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only 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-senstive 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.

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (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 http://mlg.ulb.ac.be/BruFence and http://mlg.ulb.ac.be/ARTML

Please cite: 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

Database released under Open Database License, individual contents under Database Contents License



Time 🖋	V1 🎤	V2 🌶	V3 🧳	
0	-1.3598071336738	-0.0727811733098497	2.53634673796914	
0	1.19185711131486	0.26615071205963	0.16648011335321	
1	-1.35835406159823	-1.34016307473609	1.77320934263119	
1	-0.966271711572087	-0.185226008082898	1.79299333957872	
2	-1.15823309349523	0.877736754848451	1.548717846511	
2	-0.425965884412454	0.960523044882985	1.14110934232219	
4	1.22965763450793	0.141003507049326	0.0453707735899449	
7	-0.644269442348146	1.41796354547385	1.0743803763556	
7	-0.89428608220282	0.286157196276544	-0.113192212729871	
9	-0.33826175242575	1.11959337641566	1.04436655157316	

+ Load 90 more rows

Versions			
Version 2	2 months ago	CSV format	
Version 1	2 months ago	Rdata format	

20,908 views · 3,201 downloads · 97 kernels · 10 topics

Explore More Datasets

© 2016 Kaggle Inc Our Team Careers Terms Privacy Contact/Support