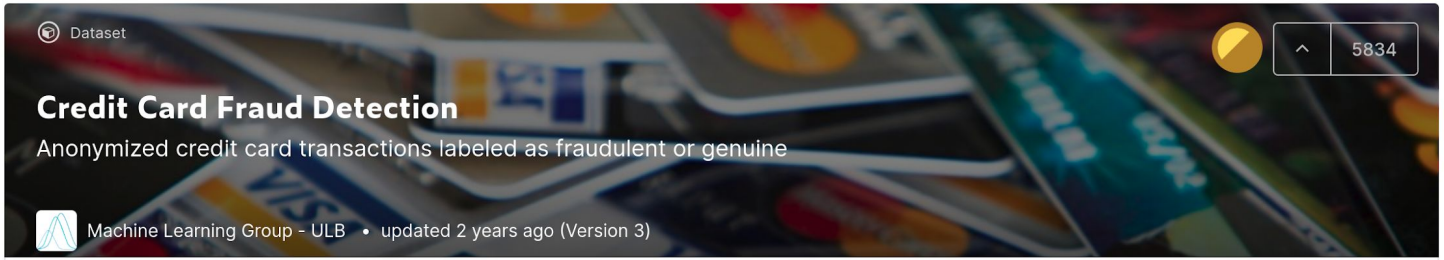
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA ĐIỆN – ĐIỆN TỬ, BỘ MÔN VIỄN THÔNG

THIẾT KẾ MÔ HÌNH PHÂN LOẠI GIAN LẬN TRONG GIAO DỊCH THẺ TÍN DỤNG

ĐỒ ÁN MÔN HỌC

Mai Văn Thời —1513319

Giáo Viên Hướng Dẫn
Thầy, Vương Phát



Data | Tasks (9) | Kernels (2,619) | Discussion (67) | Activity | Metadata | Download (144 MB) | New Notebook

Usability 8.5 | **License** Database: Open Database, Contents: Database Contents | **Tags** computer science, finance, crime

Description

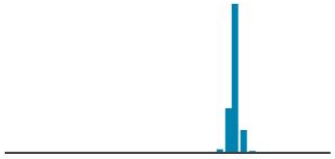
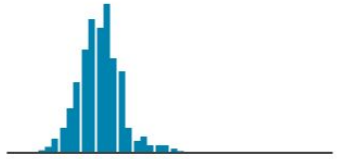
Context

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.

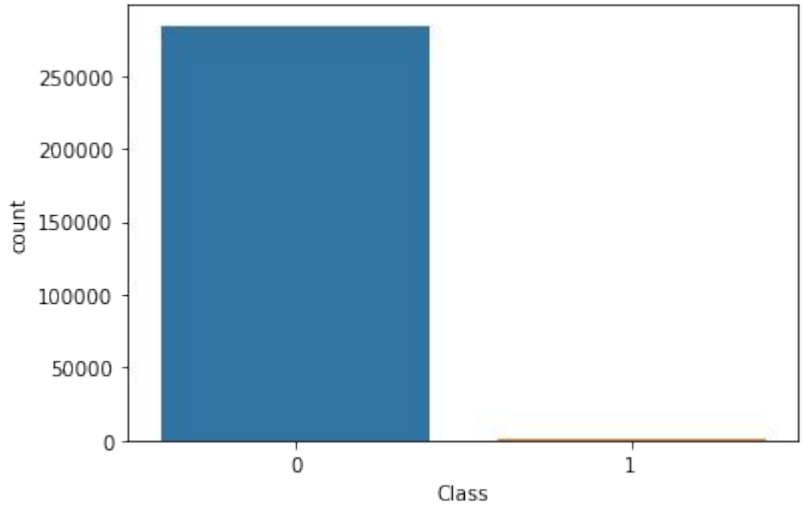
Content

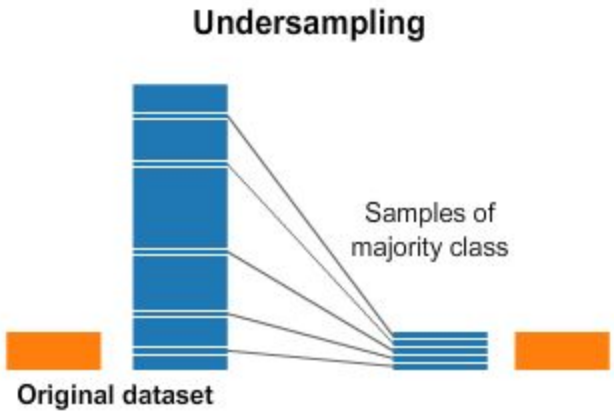
The datasets contains transactions made by credit cards in September 2013 by european cardholders.
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

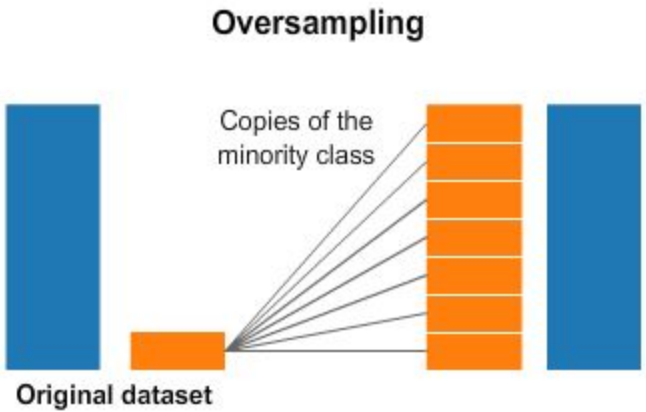
Data Explorer

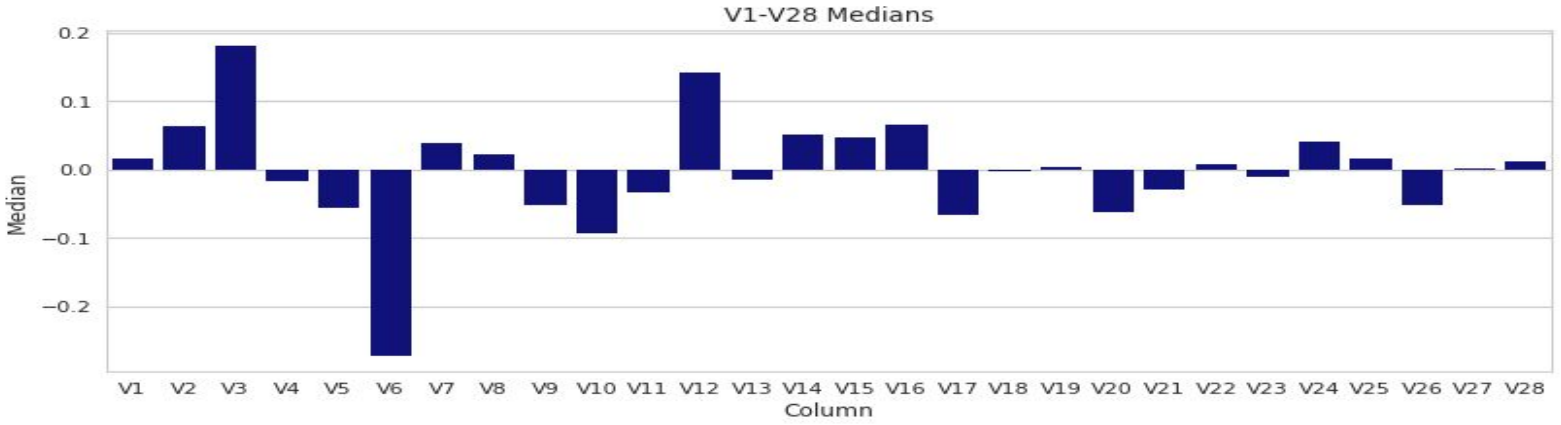
-56.4 2.45	 -72.7 22.1	-48.3 9.38	 -5.68 16.9	-114
-1.3598071336738	-0.0727811733098497	2.53634673796914	1.37815522427443	-0.33832
1.19185711131486	0.26615071205963	0.16648011335321	0.448154078460911	0.060017
-1.35835406159823	-1.34016307473609	1.77320934263119	0.379779593034328	-0.50319
-0.966271711572087	-0.185226008082898	1.79299333957872	-0.863291275036453	-0.01030
-1.15823309349523	0.877736754848451	1.548717846511	0.403033933955121	-0.40719
-0.425965884412454	0.960523044882985	1.14110934232219	-0.168252079760302	0.420986
1.22965763450793	0.141003507049326	0.0453707735899449	1.20261273673594	0.191880
-0.644269442348146	1.41796354547385	1.0743803763556	-0.492199018495015	0.948934
-0.89428608220282	0.286157196276544	-0.113192212729871	-0.271526130088604	2.669598
-0.33826175242575	1.11959337641566	1.04436655157316	-0.222187276738296	0.499360
1.44904378114715	-1.17633882535966	0.913859832832795	-1.37566665499943	-1.97138
0.38497821518095	0.616109459176472	-0.874299702595052	-0.0940186259679115	2.924584

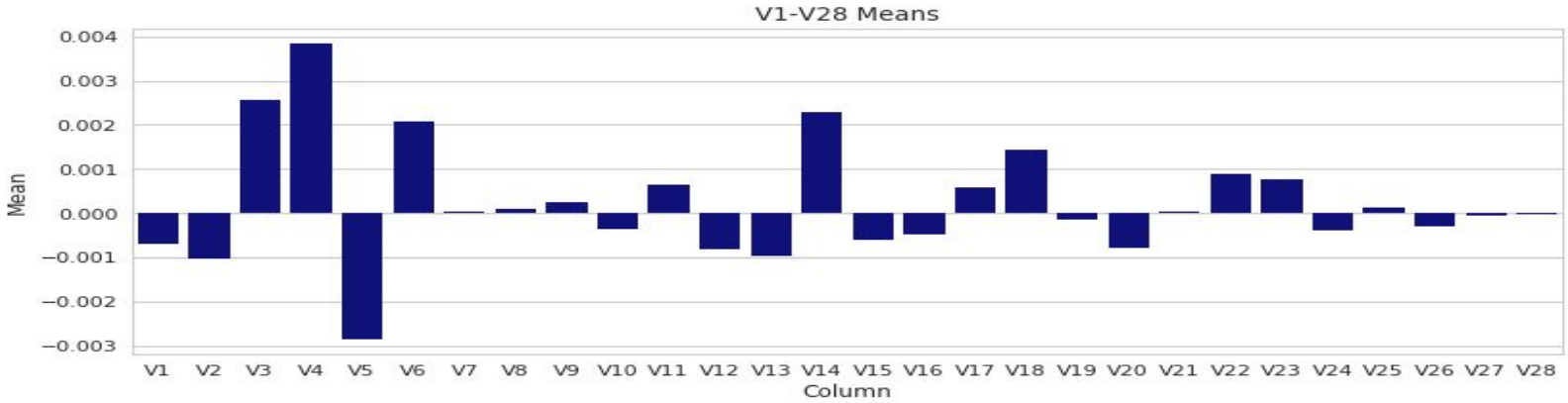
Skewed data

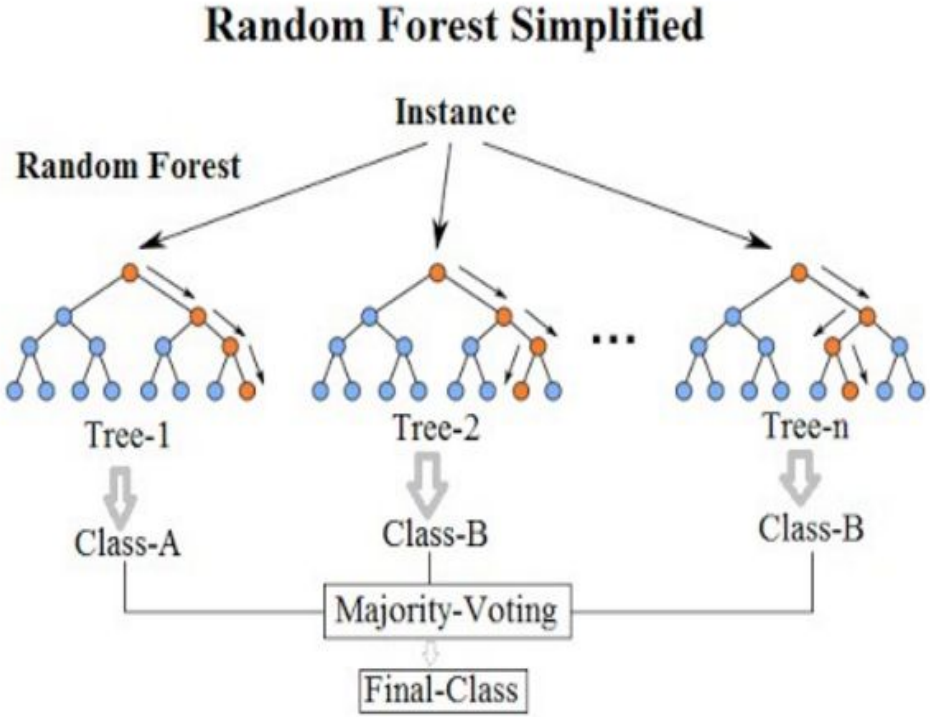


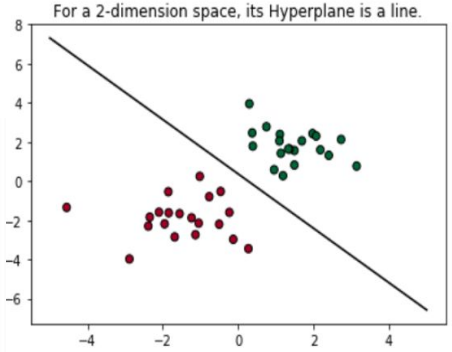

Undersampling
Samples of
majority class
Original dataset

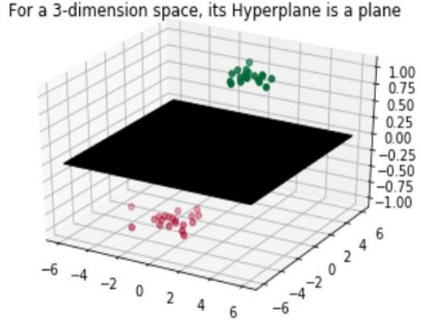

Oversampling
Copies of the
minority class
Original dataset


V1-V28 Medians
Median
0.2
0.1
0.0
−0.1
−0.2
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28
Column

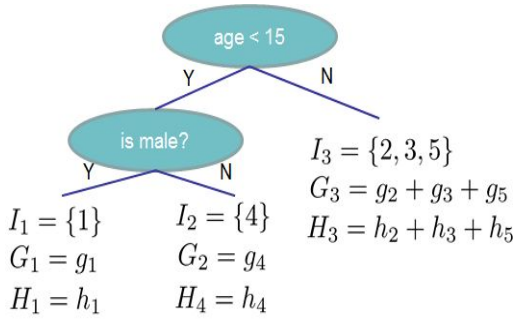

V1-V28 Means
Mean
0.004
0.003
0.002
0.001
0.000
−0.001
−0.002
−0.003
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28
Column







Instance index	gradient statistics
1	g1, h1
2	g2, h2
3	g3, h3
4	g4, h4
5	g5, h5



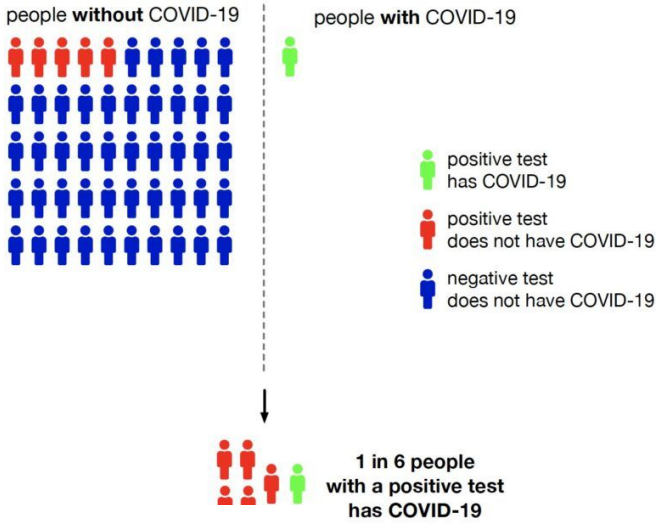
$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

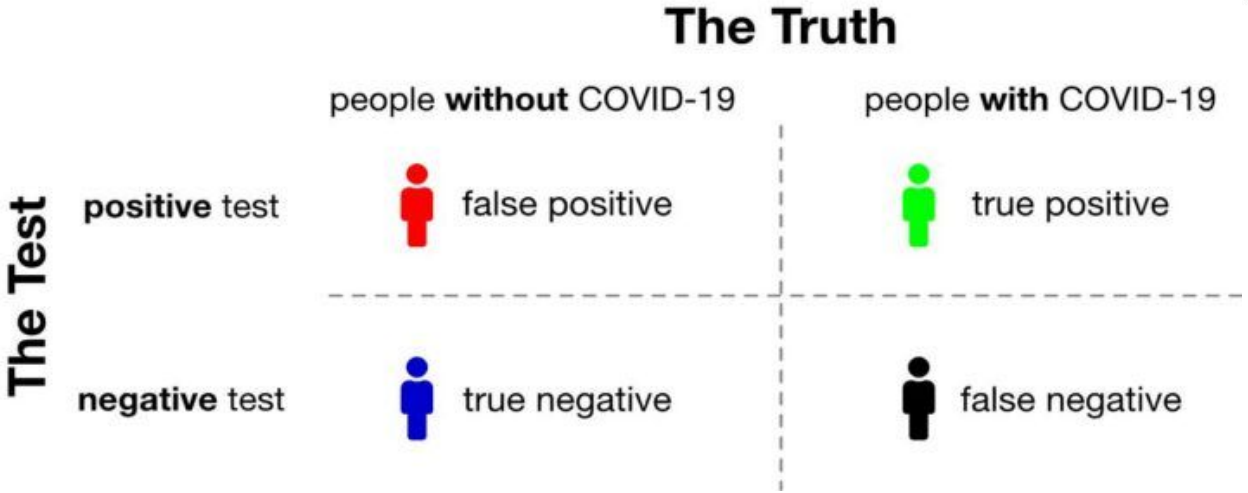
The smaller the score is, the better the structure is

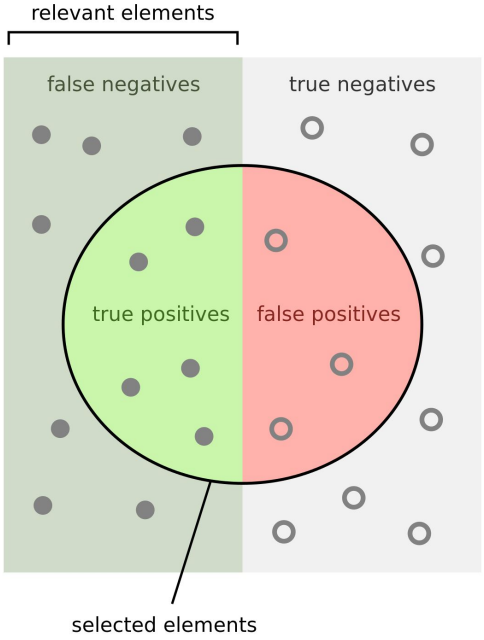
Table 1 Bảng confusion matrix chuẩn

	Predicted positive	Predicted negative
Actual positive	True positive TP	False negative FN
Actual negative	False positive FP	True negative TN

Covid19-FN >> FP
Satisfation- FP matters
Margin - FN matters

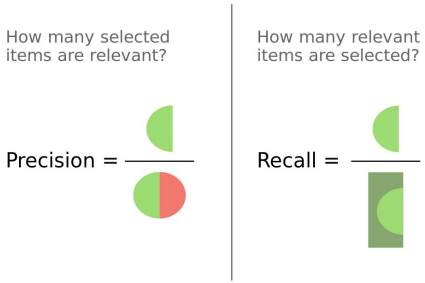






$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{n^+ + n^-} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$F_1 \text{ score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \tag{2}$$

Table 4 Bảng ranking dự đoán

Classifier	MCC	F1-Score	Accuracy
MCC ranking			
Random Forest	0.78872		
SVM	0.78626		
XGBoost	0.77352		
F1-score ranking			
SVM		0.77966	
Random Forest		0.77500	
XGBoost		0.76647	
Accuracy ranking			
Random Forest			0.99951
SVM			0.99933
XGBoost			0.99931

“It is the province of knowledge to speak, and it is the privilege of wisdom to listen.” —*Oliver Wendell Holmes Sr.*