RISK ASSESMENT

Nicolas Bens
Saúl Pérez-Silguero

GOAL

Predict wether a client will commit fraud or not

↓

CLASSIFICATION CASE

About the Dataset

Variables:

- Y: Target (fraud, no fraud)
- X: Independent variables (gender, loan, income...)

Cleaning :

- Null values
- Blank cells

Obstacle:

- Unbalanced

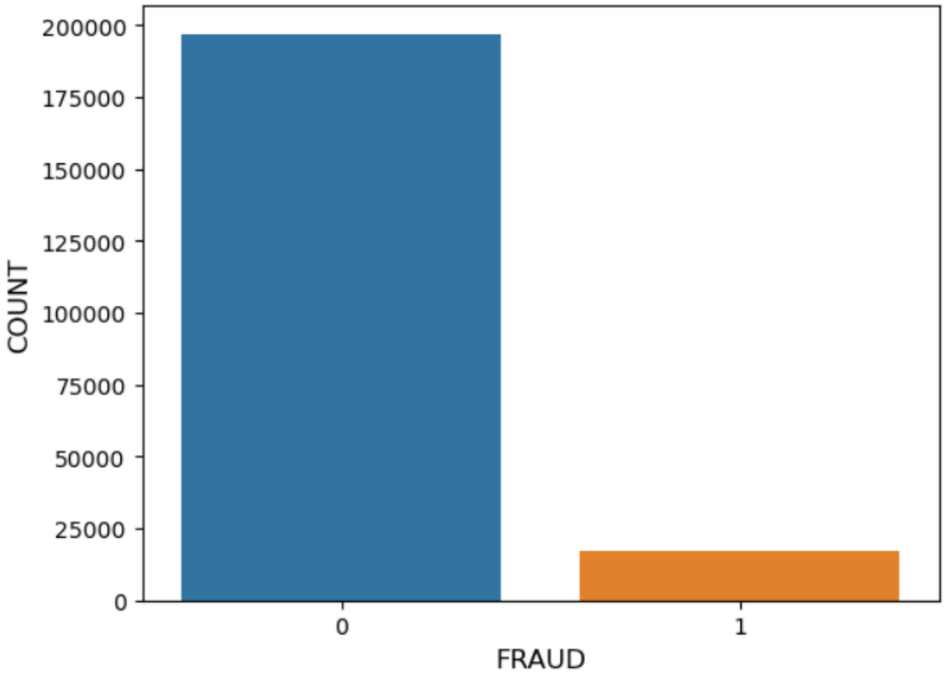
Target variable:
Not enaugh fraud cases

How can we deal with unbalanced Datasets?

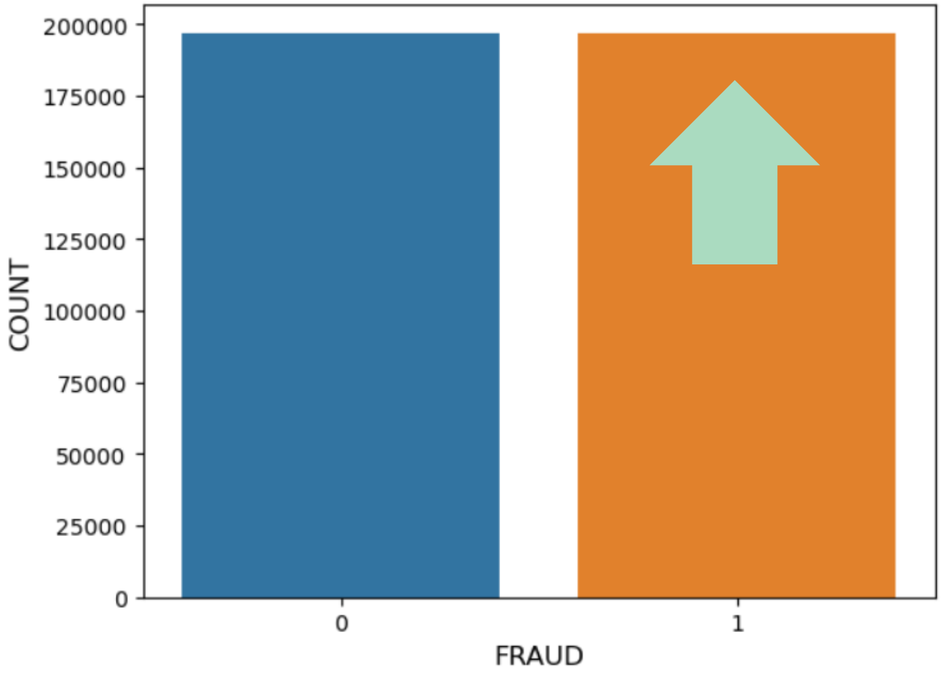
Take a chunk off the Dataset (Old-fashioned way)

SMOTE library (Cool way)

Before SMOTE



Afer SMOTE



Multiplies closest values

How can we asses the performance of a classification model?

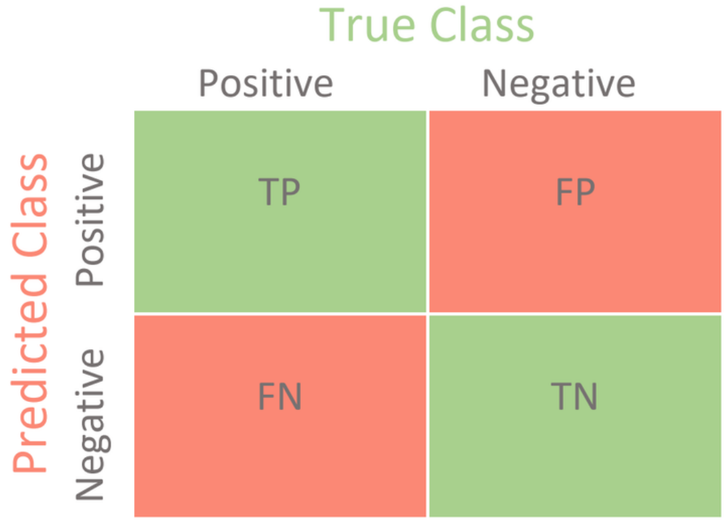
↓

Confussion Matrix

CONFUSSION MATRIX

Useful measure/table for classification cases.

Summarizes the counts of true positive, true negative, false positive, and false negative predictions.



Why is the Confussion Matrix useful?

PARAMETERS

They asses the performarce of a classification model based on the confussion matrix:

PRECISION
TP/(TP+FP)

RECALL
TP/(TP+FN)

ACCURACY
TP+TN/(TP+FP+TN+FN)

- Individually lack a comprehensive assessment of model performance.
- Python offers a function that calculates these parameters.

Parameters for each model

Chunk off and Logistic Regression

We want to predict FRAUD and aim to have a high RECALL

	precision	recall	f1-score
0	0.60	0.60	0.60
1	0.58	0.58	0.58
accuracy			0.59
macro avg	0.59	0.59	0.59
weighted avg	0.59	0.59	0.59

SMOTE and Logistic Regression

	precision	recall	f1-score
0	0.94	0.51	0.66
1	0.10	0.60	0.17
accuracy			0.52
macro avg	0.52	0.55	0.41
weighted avg	0.87	0.52	0.62

SMOTE and K Nearest Neighbours

	precision	recall	f1-score
0	0.93	0.77	0.84
1	0.11	0.32	0.16
accuracy			0.74
macro avg	0.52	0.55	0.50
weighted avg	0.86	0.74	0.79

In a nutshell

SMOTE and KNN

HIGH Accuracy
LOW Recall

SMOTE and Logistic Regression

MODERATE Accuracy
HIGH Recall

Chunk off and Logistic Regression

MODERATE Accuracy
HIGH Recall


Logistic
Regression
is the best
model

Thank you!