Data challenge:
Non Performing Loans

Alberto Remoto

Overview

- Non-Performing Loans are an inherent part of a bank's activities

- Fluctuations in the NPL over the course of an economic cycles are common and not worrisome as long as they remain within reasonable proportion

- Above a certain level, NPLs hamper the lending capabilities of the banks and impact economic growth

- As of Q3-2016, NPLs of significant institutions in the Euro area amounted to €921bln (average NPE of 7%)

- As per request of the ECB, it become important to set up strategies to manage and reduce NPL ratio

Challenge objective

- The data challenge requires to define a strategy **to re-rank NPL customers according to their likelihood** to repay and/or by **suggesting new strategies to maximise the recovery rate**

- The dataset provided consists of a series of CSV files that contains synthetic but yet realistic data on NPL

- The data analysed cover the historical period from Q1-2015 to Q3-2017 (33 months) and contains 27’888 unique NPLs

Strategy

1. Estimate the likelihood to repay, i.e. **probability that a payment is performed**

2. Estimate the recovery rate, i.e. **amount recovered w.r.t. the total debt**

- Both estimates will be given within a time window of 12 months

- A supervised machine learning approach seems appropriate:

    1. A classifier trained to learn if an NPL is repaid/not-repaid provide an estimate of the likelihood to repay

    2. A regressor trained to learn the realised recovery rate on the available historical dataset will provide an estimate of the recovery rate

Target variables

- The supervised machine learning approach requires to define target variables to be used in the model training:

- For the regression task:

  - The target variable is defined by the sum of amount (*incassi*) received in the next 12 months, divided by the GBV of the debt at that given time

$$Target(t) = \frac{\sum_{i=t}^{t+12} Incassi(i)}{GBV(t)}$$

- For the classification task:

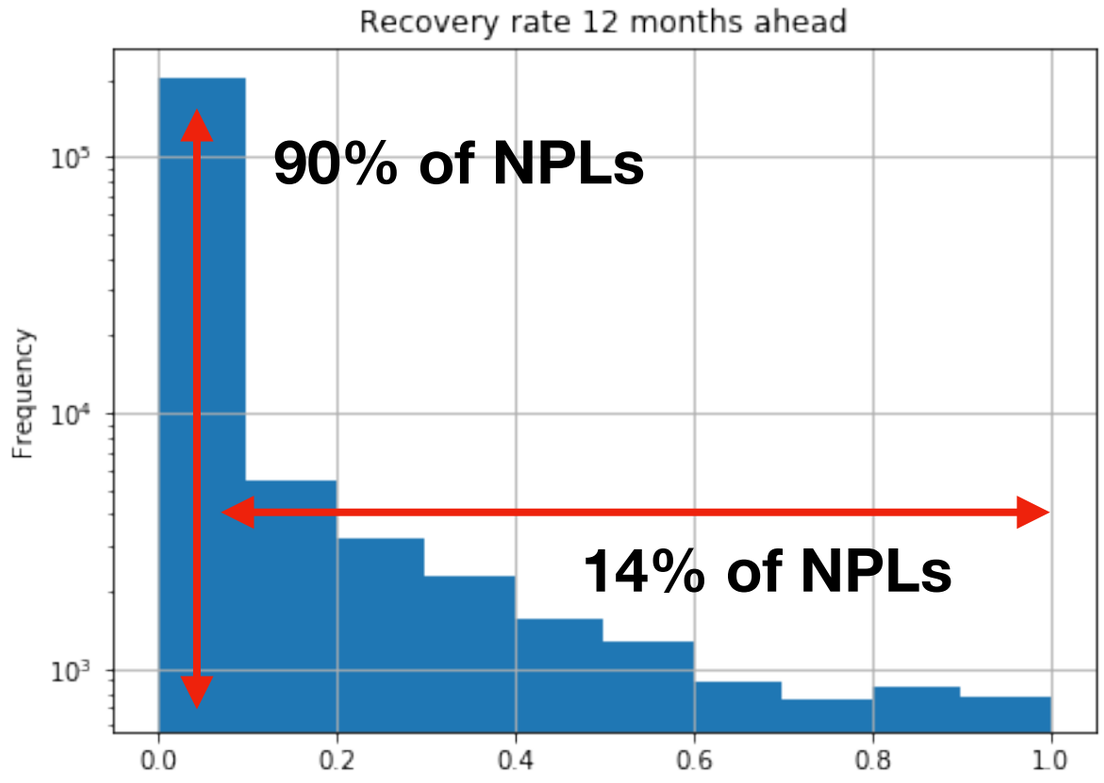
  - The target variable is defined as a simple True/False to indicate if any income exists in the following 12 months

Target variables

- The plot show the distribution of the target variable which represent the recovery rate 12 months ahead

- What is clear is that for more than 90% of the NPL, the recovery rate is < 10%

- As a matter of fact, historically only 14% of NPL provide some recovery.



Features selection

- Features selection was performed considering only those features that provide some predictive power with respect to the target variable (Pearson's correlation > 5%)
- Based on a measurable and reproducible approach
- Avoid *black-box* approach were features are automatically selected by the machine learning model

Features selection

#	Feature	Description	Correlation*
0	IMP_GBV	Total debt gross book value	7.4 %
1	IMP_CUSUM_INCASSI	Historical cumulative sum of paid amounts	6.4%
2	MAX_IMP_GBV	Max historical value of the gross book value	6.6%
3	RECOVERY_RATE_TOTAL	Total historical recovery rate	6.0%
4	NUM_MONTH_SOFFERENZA	Age of the NPL	-12.5%
5	CONTABILE_LINEA_INTERESSI_DI_MORA	Amount of the credit line interesrt	7.3%
6	CREDITO_VANTATO	Amount of the credit line	7.0%
7	NUM_ENTI_SEGNALANTI	Number of banks that reported the same customer	-5.0%
8	NUM_COMPONENTI_COINTESTAZIONE	Number of components in a joint account	8.0%
9	NUM_ID_RAPPORTO	Number of open accounts	5.2%
10	NUM_ID_GARANTE	Number of guarantors	7.6%
11	IMP_INCASSI_ROLLING_12M	Total paid amount 12 month ahead	6.4%
12	IMP_GAR	Total guaranteed amount	11.2%**

**(min correlation among the two target features)*
***(Only for customer type Persona Fisica)*

Best models

- Different types of model are tested on the whole dataset and the most performant model is retained: Random Forest

- The random forest is an ensemble model consisting of many decisions trees. It uses sample and feature sub-sampling when building each individual tree to create an uncorrelated forest of trees.

- The prediction of the ensemble is given by the average prediction of each tree. The ensemble prediction is more accurate and less prone to overfit (lower variance) than the one from single trees.

Models improvement

- Divide the type of customers in three different categories: *Persona Fisica*, *Persona Giuridica* and *Cointestazione*.

- Different customer types are expected to undertake different types of loans, are exposed to different types of risks and are driven by different idiosyncratic dynamics.

- Leveraging additional intuitions matured on route, I improved the the modes considering customer types separately.

- The Pearson's correlation showed that an additional feature ('IMP_GAR') should be considered for *Persona Fisica.*

Models performance

Customer type	Classifier				Regressor	
	F1	Precision	Recall	AUC	R2	MSE
All customers	0.85 (0.91)	0.89 (0.93)	0.81 (0.89)	0.98	0.66 (0.78)	0.005 (0.003)
Persona Fisica	0.90 (0.91)	0.91 (0.93)	0.88 (0.94)	0.99	0.73 (0.83)	0.004 (0.002)
Persona Giuridica	0.87 (0.93)	0.91 (0.94)	0.84 (0.92)	0.99	0.72 (0.83)	0.004 (0.002)
Cointestazione	0.89 (0.93)	0.91 (0.94)	0.86 (0.92)	0.98	0.70 (0.82)	0.005 (0.003)

Precision: proportion of correctly predicted instances

Recall: proportion of actual prediction that are correctly identified

F1: harmonic mean of Precision and Recall

AUC: measure the probability the classifier perform better than a random choice

R^2: proportion of variance explained by the model

MSE: measures the average of the squares of the errors.

Training separated models for different customer types improve the performance by 3-5% (F1) for the classification task and by 5-7% for the regression task.

Predict and rank NPLs

- I apply the predictive models over the most recent NPLs (09-2017) in order to estimate the **probability of recovery** and the **expected recovery rate** 12 months ahead.

- The list of NPLs is then ranked by those two predicted values and made available on the *npl_sorted.csv* file.

- In order to facilitate the work of portfolio managers, the probability of recovery is used to define a **recovery class**, i.e. group of NPLs within a given range of probability.

Portfolio composition and expected recovery rate

- **Recovery Class A** contains the NPLs which are more likely to recover (Prob > 75%). A total of 1'427 NPLs (7.6%) are in this class. The total debt corresponding to this class is 7.5M€, the predicted recovery rate is 10% which correspond to recover about 0.8M€.

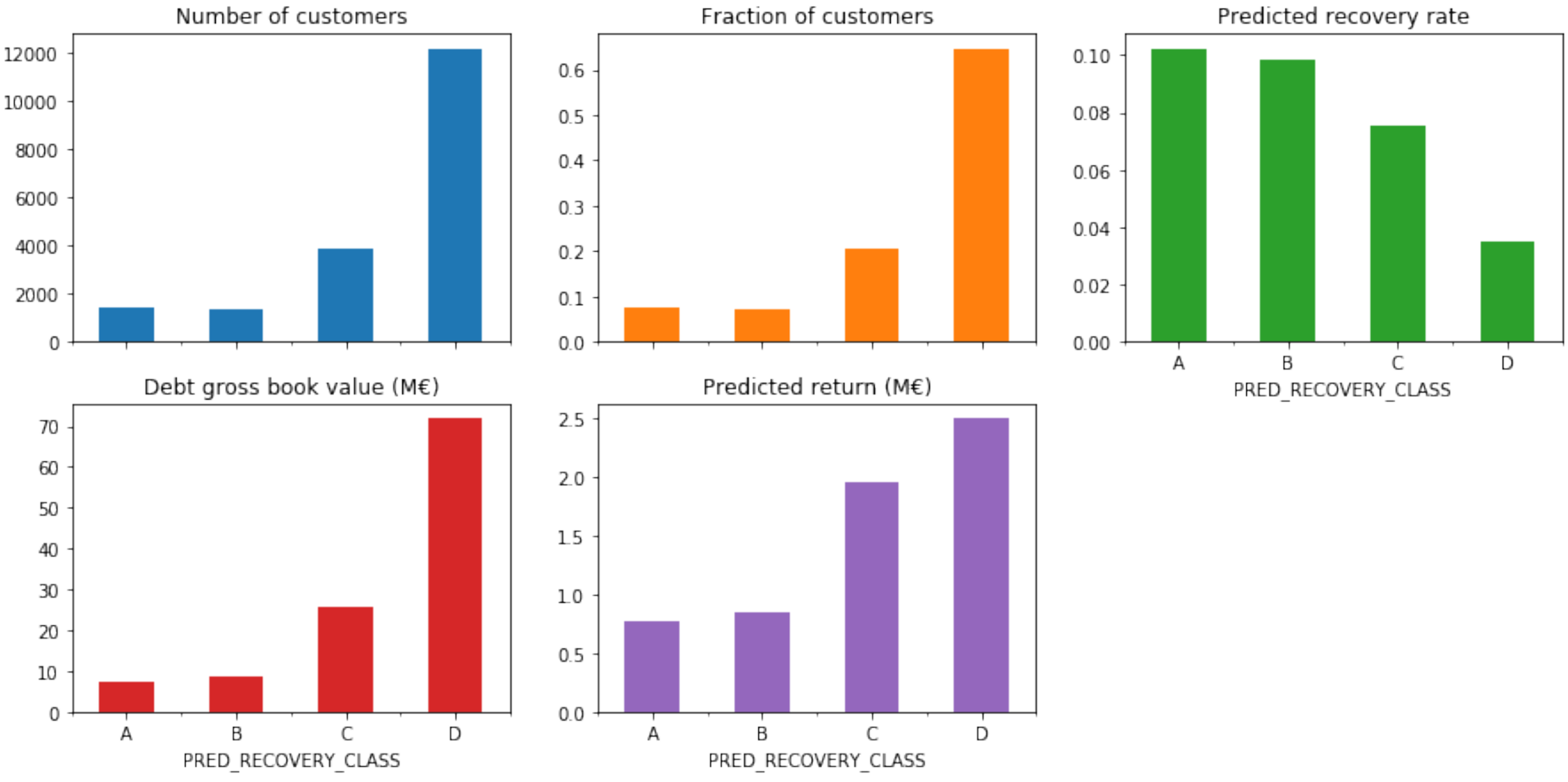
- **Recovery Class B** contains the NPLs which are likely to recover (Prob within 50% and 75%). A total of 135'727 NPLs (7.2%) are in this class. The total debt corresponding to this class is 8.6M€, the predicted recovery rate is 10% which correspond to recover about 0.8M€.

- **Recovery Class C** contains the NPLs which are not likely tro recover (Prob within 25% and 50%). A total of 3847 NPLs (20%) are in this class. The total debt corresponding to this class is 25.8M€, the predicted recovery rate is about 7.5% which correspond to recover about 1.9M€.

- **Recovery Class D** contains the NPLs which are very unlikely to recover (Prob < 25%). A total of 12'153 NPLs (64%) are in this class. The total debt corresponding to this class is about 72M€, the predicted recovery rate is 3.5% which correspond to recover about 2.5M€.

- **Given the NPL sample as of 09-2017, the expected recovery rate that is between 1.6M€ (only A and B classes) and 6M€ (also C and D classes).**

Portfolio composition and expected recovery rate



Recommendation for NPL portfolio managers:

Keep NPLs type A,B & C in the portfolio. Take NPLs type D Off the Book

Business case

XXX = Estimates, XXX = Computed

Year	# NPL	GBV (M€)	%	Incassi (M€)	# OTB	OTB (M€)	# New	% A	% B	% C	% D	Cost (€)
2015	16 137	78		0,0	544	2,1		8,5	7,5	23	61	-
2016	23 428	106	36	4,5	5 156	18,9	8 642	10	7	22	61	-
2017	20 984	114	8	5,0	2 122	9,1	2 606	9	12	36	43	-
F2018	11 484	55	10	3,6	12 000	70	2 500	10	10	20	60	63092
F2019	12 484	52	10	3,6	1 500	9	2 500	10	10	20	60	14092
F2020	13 484	49	10	3,6	1 500	9	2 500	10	10	20	60	14092
		Totals (M€)		**10,8**		**88**				**Total cost (€)**		**91275**

Costs Breakdown

Resource		Cost per unit (€)	Unit	N. Units	Cost (€)	Notes
Data Engineer		70000	FTE/y	0,1	7000	1 Data Engineer, 1 month full time for infrastructure deployment.
Software Developer		70000	FTE/y	0,5	35000	1 Developer, 6 month full time for client UI development.
Data Scientist		70000	FTE/y	0,2	14000	1 Data Scientist, 2 day a week for model maintenance.
Data Storage		0,022	€/GB/m	120	2,64	AWS S3 service up to 50 TB, or equivalent. 10 GB/month x 12 months.
Production environment	**Deployment**	0,57	€/h	52	29,64	AWS SageMaker service, or equivalent. Consider ml.m5.2xlarge instance (8 CPU cores, 32 GB RAM, elevate network performance) for model deployment, training and real time model inference. 1h/week x 52 weeks for each step.
	Training	0,57	€/h	52	29,64	
	Inference	0,57	€/h	52	29,64	

Summary

- The models relies on 13 features chosen by their strong predictive power. Models' performances are quite good and provide reliable ranking of NPLs.

- Within Agile methodology, this is the first release for production. Future improvement to be expected, based on additional finding & feedbacks.

- Many possible paths to be explored for improvement, for example:
    1. Include features excluded so far;
    2. Test extensively with different combination of features;
    3. Train different models for different geographical areas;

- If implemented, in 3 years this solution is expected to reduce the NPL GBV by 57%, to collect about 11M€ from existing NPL, to take about 88 M€ off the books. The estimated cost of the solution is about 100k€.

Backups

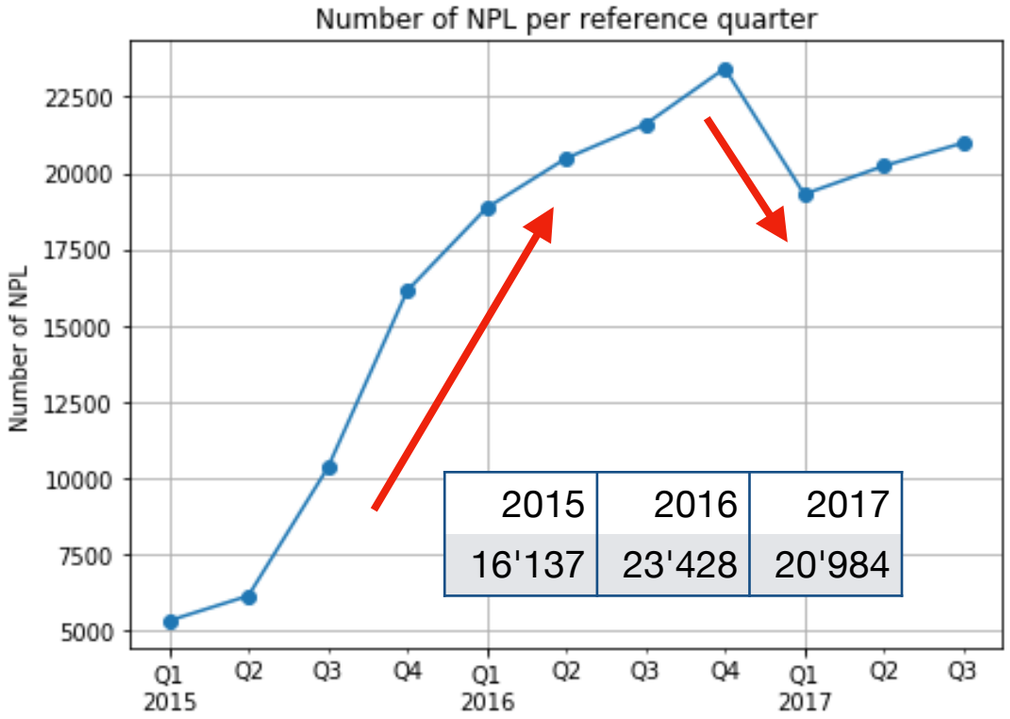
Data overview (1)

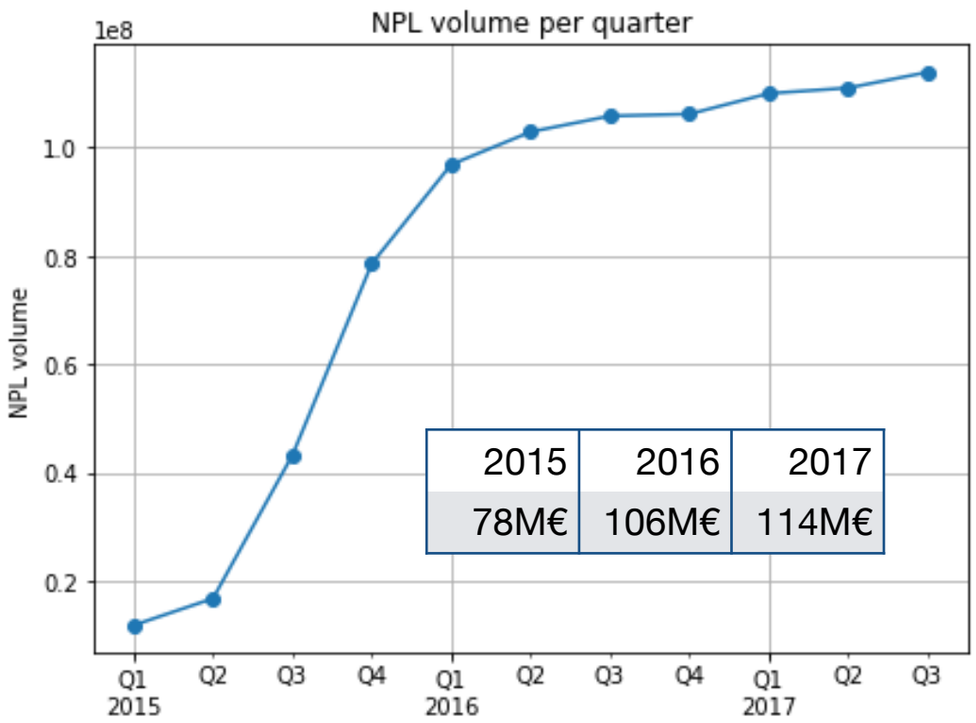
- The data cover the historical period from Q1-2015 to Q3-2017 (33 months) and contains 27’888 unique NPLs

- The number of NPL increase from Q1-2015 even though the grow slowed down in Q1-2016

- The total NPL volume follow a similar trend, as expected

- The number of NPL reduce in Q1-2017 as a consequence of action taken to reduce NPE





Data overview (2)

- The average recovery rate is about 4% of the total GBV a year

- About 18% of the GBV is taken off the book in 2016 as a measure to reduce NPE

- This actions reduce the recovery rate in Q4-2016 by more than 30%

- An additional 8% of the GBV is taken off the book in 2017 but this time the recovery rate remained unchanged