ITA DATA 2024

ITADATAhack

PPG

17/09/2024

Con il patrocinio di

Outline

1 - Our Team

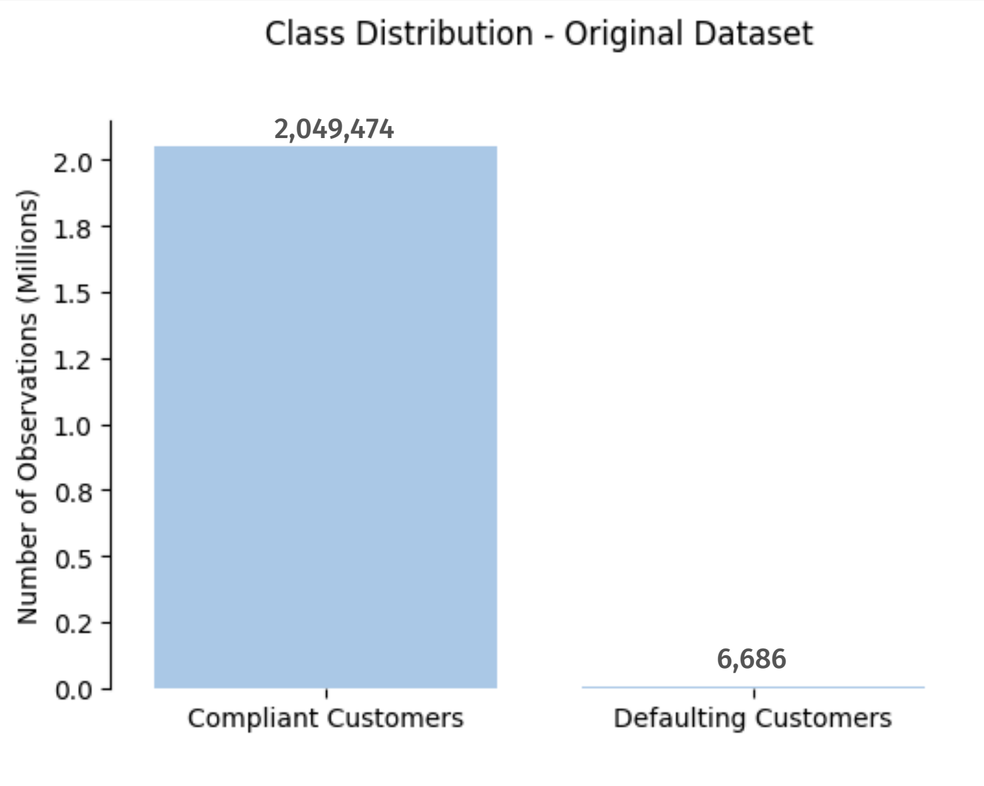
Giordano Vitale

Pietro Padovese

Jan Philip Richter

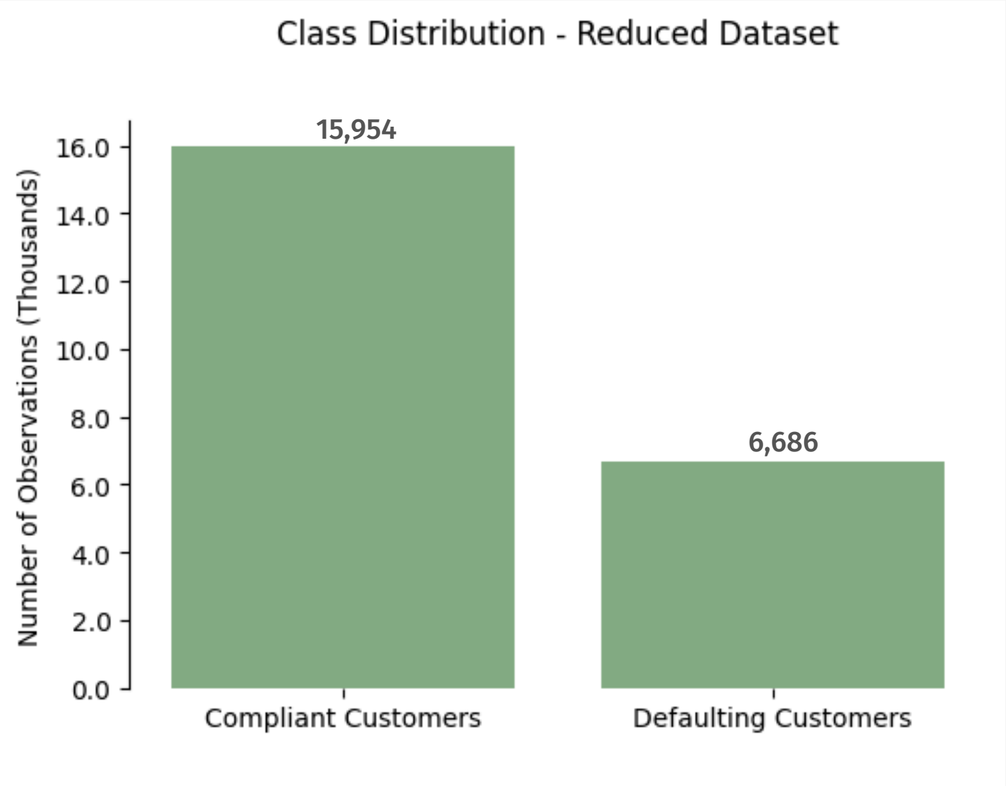
2 - Common Insights

Huge and unbalanced original dataset: 99.67% observations are compliant customer, whereas only 0.33% observations presented credit unworthiness.



2 - Common Insights

Our strategy to face the problem: **custom subsampling**. What happens if we only retain the customers that had at least once been classified as defaulting? Only 1,132 clients kept: 22,640 observations.



2.1 - Data structuring & Feature Engineering

When working with **panel data**, a crucial source of information is derived from variables from **previous periods**. Therefore, it was necessary to reorganize the data so that each row corresponds to an observation containing **all relevant information**.

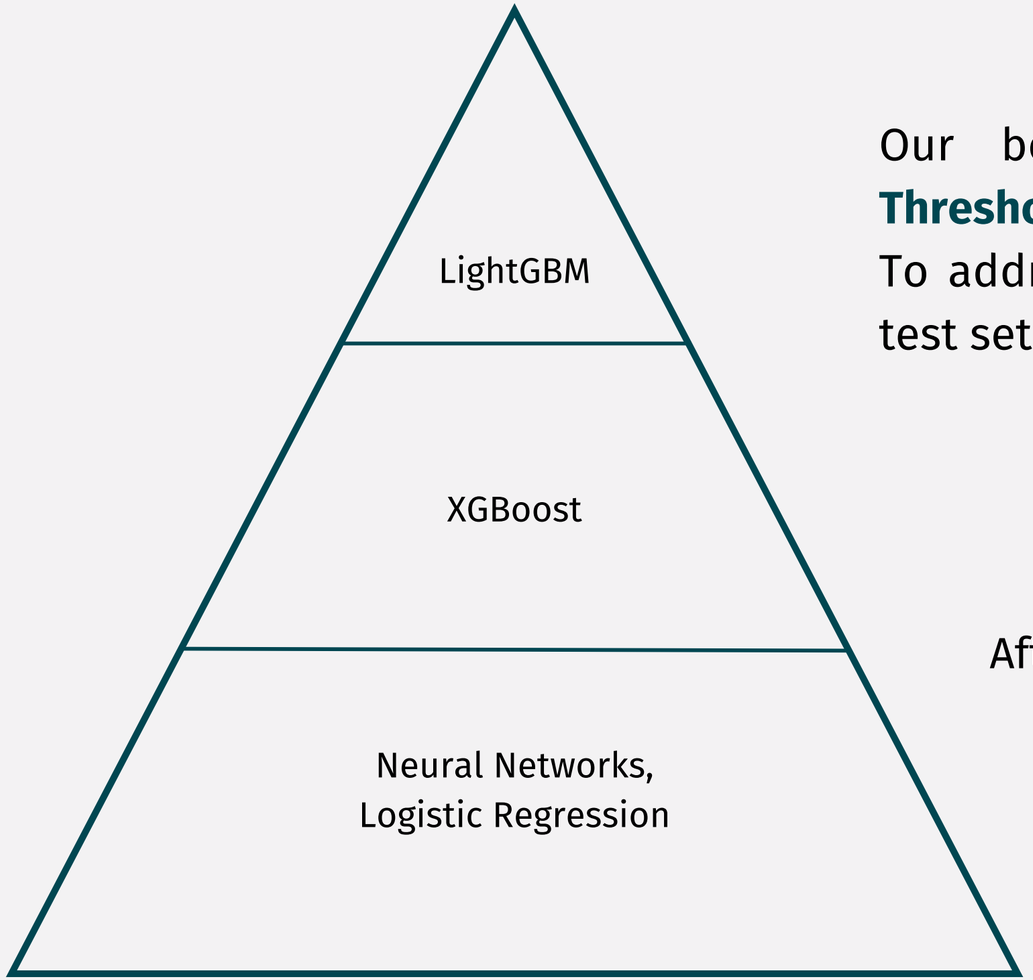
$Client_{id}$	$Period$	X_t
1	1	$x_{1,1}$
1	2	$x_{1,2}$
...	...	...
i	$T-1$	$x_{i,T-1}$
i	T	$x_{i,T}$

$Client_{id}$	$Period$	X_t	X_{t-1}
1	2	$x_{1,2}$	$x_{1,1}$
1	3	$x_{1,3}$	$x_{1,2}$
...	...	...	...
i	$T-1$	$x_{i,T-1}$	$x_{i,T-2}$
i	T	$x_{i,T}$	$x_{i,T-1}$

3) TASK 1

One-step ahead forecast

3.1 - Models



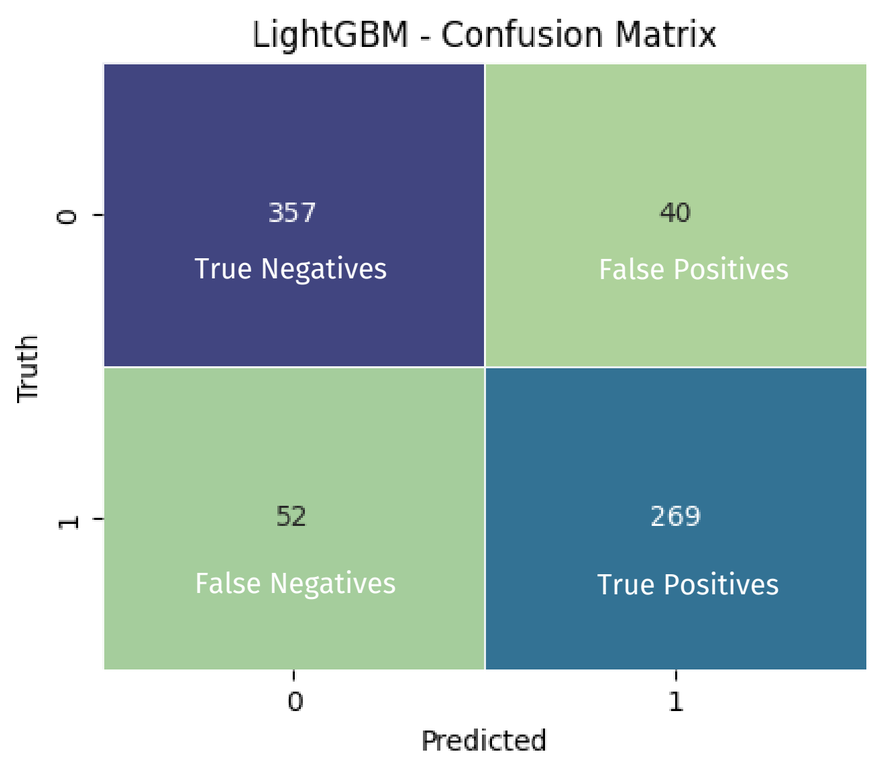
Our best performing model: **LightGBM + Custom Decision Thresholding**

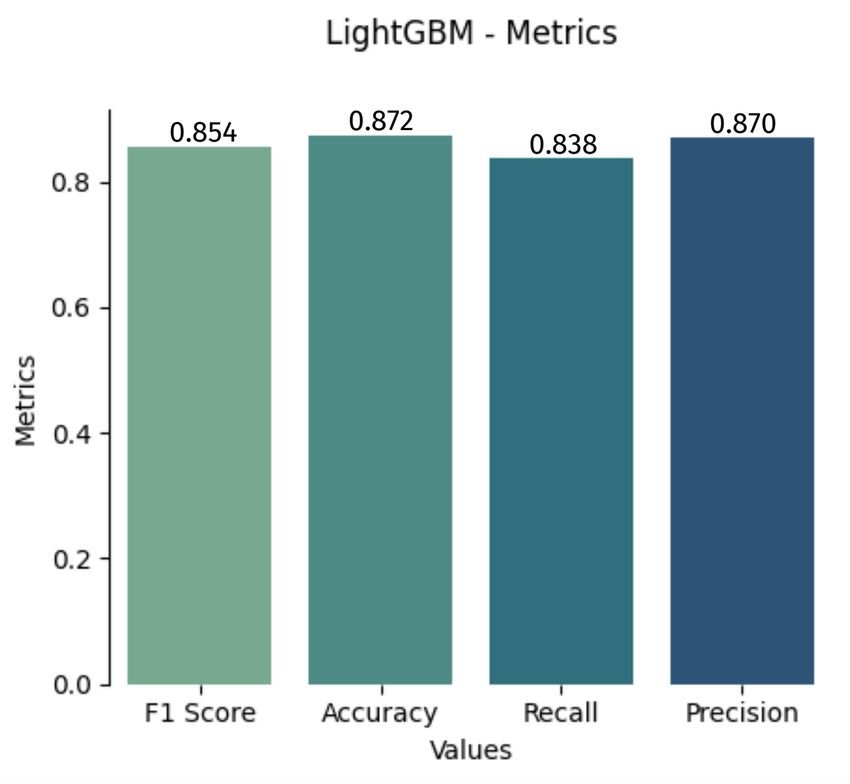
To address the remaining imbalance, the predicted values for the test set were tranlsated into binary predictions as follows:

$$\hat{y} = \begin{cases} 0 & \text{if pred\_proba} < t \\ 1 & \text{otherwise} \end{cases}$$

After several tweakings, the optimal t*=0.985

3.2 - Model Results



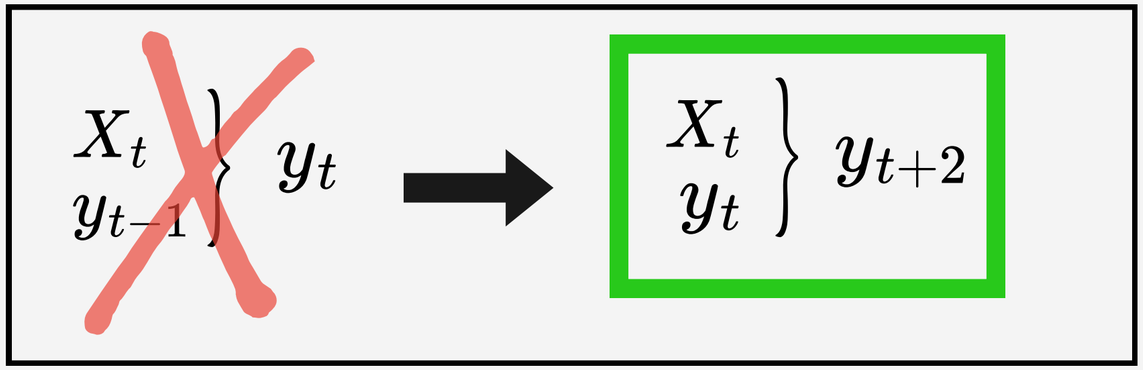


4) TASK 2

Longer forecast horizon!

4.1 - The new problem

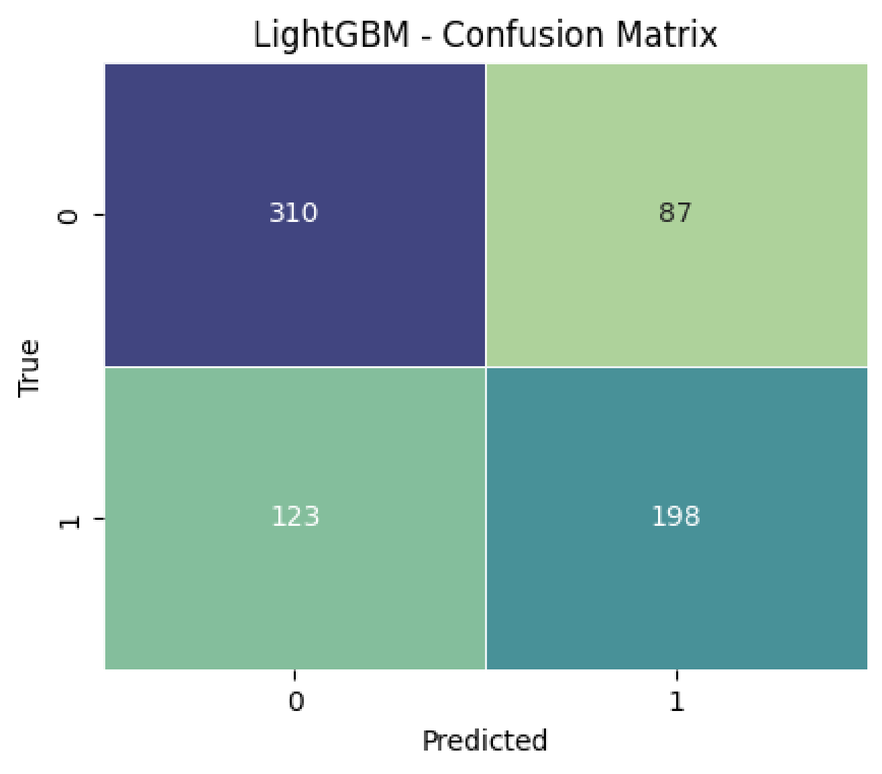
As we extended the prediction horizon, our initial thought was to **repurpose** the model created for Task 1, substituting the missing data with **alternative predictions.**

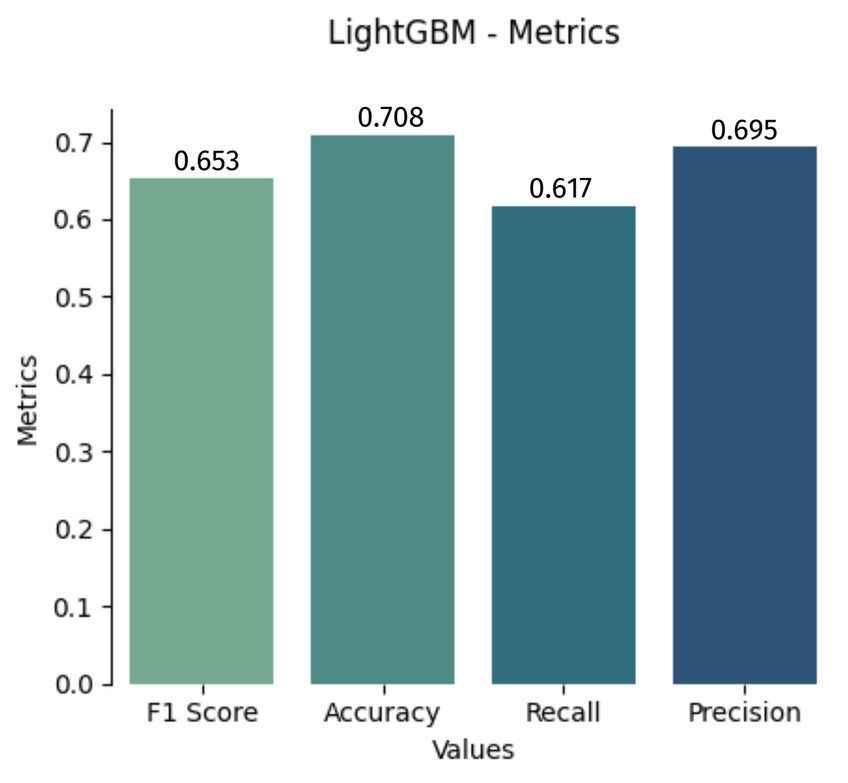


Period	X_t	X_{t-1}	y_t	y_{t-1}
T	x_T	x_{T+1}	y_T	y_{T-1}
$T+1$	$\hat{x}_{T+1}$	x_T	$\hat{y}_{T+1}$	y_T
$T+2$	$\hat{x}_{T+2}$	$\hat{x}_{T+1}$	$\hat{y}_{T+2}$	$\hat{y}_{T+1}$

4.2 - Results & Performance

In the end we didn’t manage to put into practice our idea, and we decide to try the model that we used before and just adding the new infromation yt.

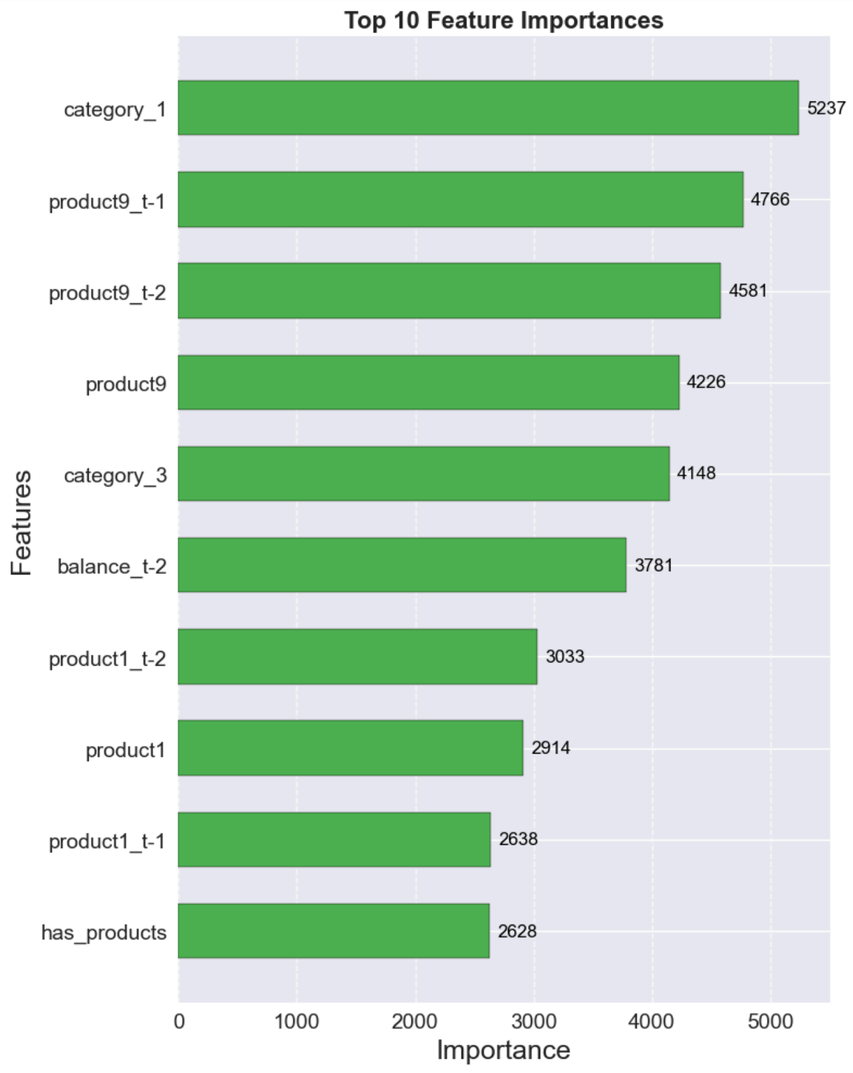




5) Feature Importance

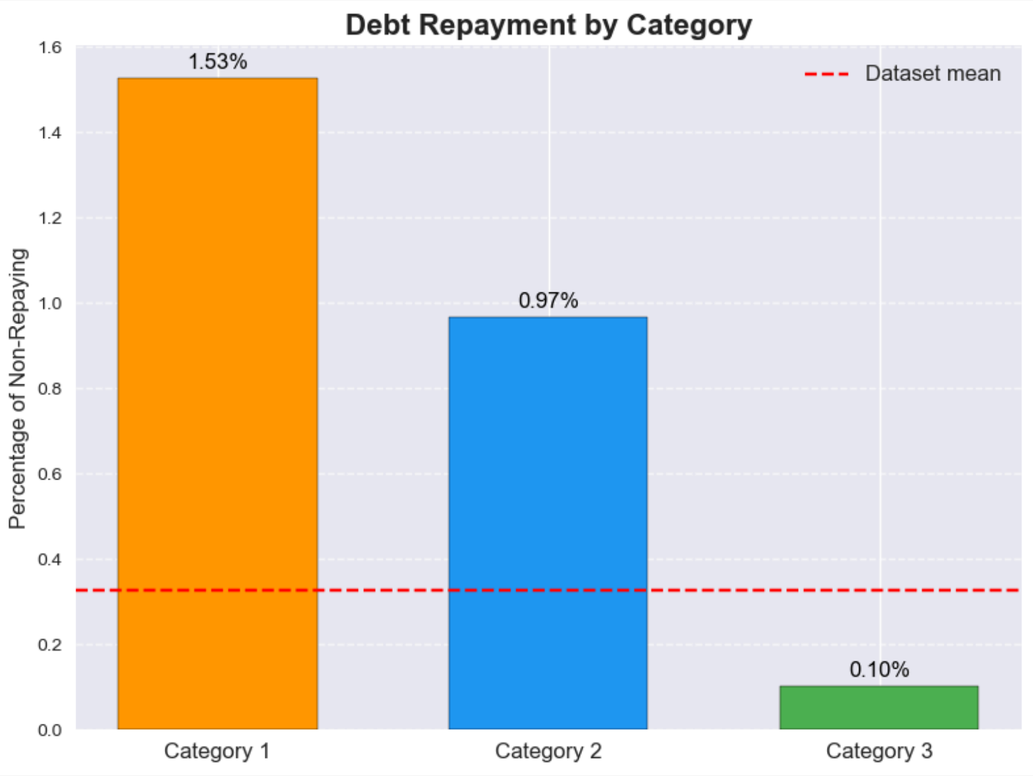
Interpret the results

5.1 - Conclusions



One of the advantages of the LightGBM model is the possibility of extracting **feature importance**, which helps us understand which variables play a **crucial role** in the decision-making process.

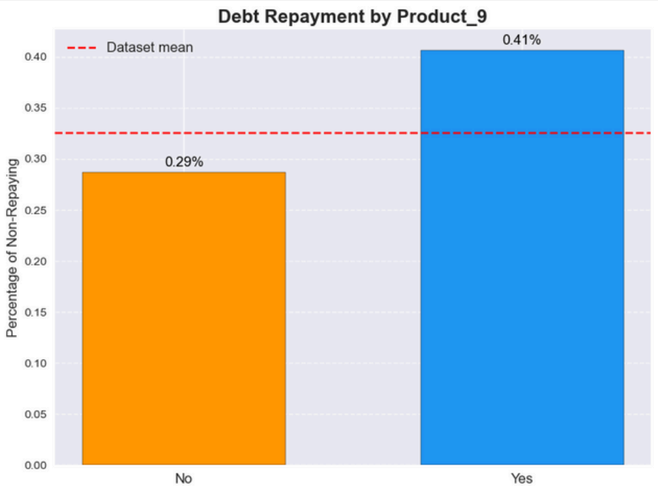
5.2 - Category

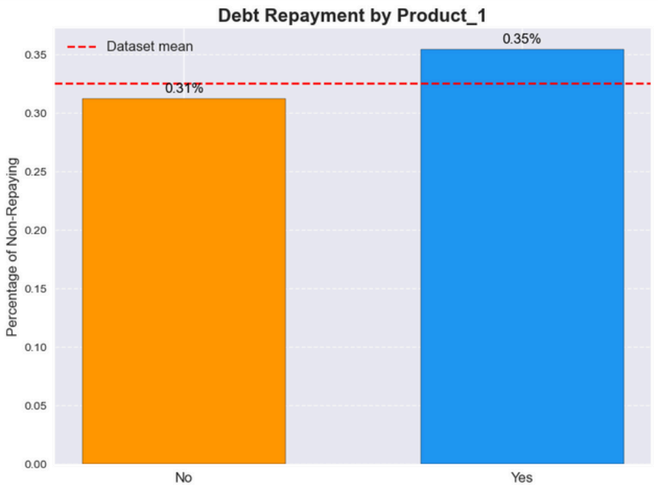


- *category_1* and *category_3* are both among the **top variables** by feature importance
- There is a **significant difference**[1] in the percentage of non-repaying customers across these categories
- This significant variation confirms that this feature is **crucial** in understanding and predicting non-repayment behavior.

1. *Proved significant by a Chi-Square test of independence*

5.3 - Product Features





- Both *product9* and *product1*, along with their corresponding lagged variables, rank among the **top 10 variables** in terms of feature importance.
- In both instances, the number of non-repaying customers is **greater** when these features are **present**.[1]
- This trend is especially notable for *product_9*, which holds a higher rank in feature importance (4226 compared to 2914). Consequently, it deserves increased focus when predicting non-repaying customers moving forward.

1. *For both variables the mean difference is proved significant by a Chi-Square test for independence*

Feel free to ask questions!

Thanks for your attention!

Organizers

opendata PLAYGROUND

DATA SCIENCE | CINI NATIONAL LAB

Consiglio Nazionale delle Ricerche

Università degli Studi di Napoli Parthenope

Università degli Studi di Bari Aldo Moro

A.D. 1308 unipg Università degli Studi di Perugia

In Supremae Dignitatis 1343 | Università di Pisa

Technical Partner

Palantir

FOURTH AGE

HPC4AI

Data Provider

ION

Sponsor

RETEX

Educational Partner

Fastweb Digital Academy

HR Partner

OPEN SEARCH NETWORK

Sustainability Partner

RIFORESTAZIONE AD ALTO IMPATTO SOCIALE

With the patronage of

cini consorzio interuniversitario nazionale per l'informatica

REGIONE TOSCANA