Description of the problem to be addressed during the hackathon:

During this first day of the hackathon, participants will face the following challenge:

- *A binary classification problem:*
  Participants will have to develop models capable of correctly classifying companies as "default" or "non-default".

To do this, they will have a training set on which to train the models and a test set on which to evaluate the models' performance.

Dataset Description:

The datasets provided are CSV files named "train_set.CSV" and “test_set.CSV” which contain various indexes relating to some companies. The two variables that indicate whether the companies have entered default and, if so, after how many days are also inserted in the train file so as to allow the training of the developed model. Below we list the variables present in the dataset:

Here is the list of features with their descriptions in the required format:

- ~~*application_ID*: Application identifier~~
- *decision_date*: Date of the decision
- ~~*company_ID*: Company identifier~~
- [redacted]core provided by an external company regarding the goodness of the company (on the scale, 1 is better)
- ~~*external_score_ver02*~~: Score provided by an external company regarding the goodness of the company (on the scale, 1 is better)
- *late_payment_score*: Score provided by an external company regarding the goodness of the company - it is linked to delays in payments (on the scale, 0 is better)
- *external_score_late_payment_integrated*: Score that integrates the scores external_score_ver01 and late_payment_score
- *external_score_moderate*: Score provided by an external company regarding the goodness of the company - defined in a moderate future macroeconomic scenario (on the scale, 1 is better)
- *external_score_adverse*: Score provided by an external company regarding the goodness of the company - defined in an adverse future macroeconomic scenario (on the scale, 1 is better)
- *external_score_ver03*: Score provided by an external company regarding the goodness of the company (on the scale, A is worse)
- *age*: Company age (in years)
- *province*: Province
- *juridical_form*: Legal form

- *industry_sector*: Industrial sector
- *gross_margin_ratio*: The ability of a company to generate excess funds from its operations, given a certain amount of assets
- *core_income_ratio*: The ratio between the adjusted earnings before interests and taxes (EBIT) and the revenues
- *cash_asset_ratio*: A conservative measure of a company's ability to pay its short-term liabilities, as it measures how much of the current assets would be available to pay off interests on debt
- *consolidated_liabilities_ratio*: Ratio between non-current (long term) liabilities and total liabilities. It represents how much of a company's liabilities are due in the next 12 months, highlighting pressing short-term debt positions in the balance sheet
- *tangible_assets_ratio*: Portion of non-current assets represented by tangible items, such as real estate and machinery, as opposed to intellectual property and goodwill. This number provide a high-level assessment of the quality and solidity of strategic balance sheet assets
- *revenues*: Company revenues
- *cr_available*: Boolean indicating the availability of information from Centrale Rischi (Bank of Italy database)
- *region*: Region
- *geo_area*: Geographic area
- *last_statement_age*: Time elapsed (in years) since the last balance sheet was filed
- *overrun_freq_a_revoca_autoliquidanti*: Percentage of months with unpaid amounts in the last year for the "revokeable/self-liquidating" product class (or in the last observable period if a full year is not available)
- *avg_tension_a_revoca_autoliquidanti*: Average over an annual time horizon of the ratio between the used and the agreed for the "withdrawal / self-liquidating" product class (i.e. the company has agreed with the banking system to be able to take advantage of a certain amount of money, the tension indicates the percentage of money used)
- *std_tension_a_revoca_autoliquidanti*: Standard deviation over annual time horizon of the above variable
- max_tension_a_revoca_autoliquidanti: Maximum value observed over the annual time horizon of the above variable
- *last_tension_a_revoca_autoliquidanti*: Last observed value of the above variable
- *avg_rel_used_a_revoca_autoliquidanti*: Average over an annual time horizon of the ratio between used and turnover for the "withdrawal / self-liquidating" product class
- *std_rel_used_a_revoca_autoliquidanti*: Standard deviation over annual time horizon of the above variable
- *max_rel_used_a_revoca_autoliquidanti*: Maximum value observed over the annual time horizon of the above variable
- *last_rel_used_a_revoca_autoliquidanti*: Last observed value of the above variable

- *overrun_freq_a_scadenza*: Percentage of months with unpaid amounts in the last year for the "maturity" product class (or in the last observable period if a full year is not available)
- *avg_rel_used_a_scadenza*: Average over an annual time horizon of the ratio between used and turnover for the "expiring" product class
- *std_rel_used_a_scadenza*: Standard deviation over annual time horizon of the above variable
- *max_rel_used_a_scadenza*: Maximum value observed over the annual time horizon of the above variable
- *last_rel_used_a_scadenza*: Last observed value of the above variable
- *avg_count_enti_affidanti*: Average over an annual time horizon of the number of requests from entities that entrust money to the company
- *std_count_enti_affidanti*: Standard deviation over annual time horizon of the above variable
- *max_count_enti_affidanti*: Maximum value observed over the annual time horizon of the above variable
- *last_count_enti_affidanti*: Last observed value of the above variable
- *avg_count_numero_prima_info*: Average over an annual time horizon of the number of requests for credit to the banking system by the company
- *std_count_numero_prima_info*: Standard deviation over annual time horizon of the above variable
- *max_count_numero_prima_info*: Maximum value observed over the annual time horizon of the above variable
- *last_count_numero_prima_info*: Last observed value of the above variable
- *days_to_default*: Days taken by the company to go into default.
- *target*: Binary variable indicating whether the company has defaulted (1) or not (0)

Results Evaluation:

- For the binary classification problem, the metric to evaluate the performance of the models will be the F1 score, calculated as:

$$F1 = 2\times\frac{precision\times recall}{precision+recall}$$

Where:

- $Precision = (\frac{TP}{TP+FP})$, where TP are the true positives and FP are the false positives.

- $Recall = (\frac{TP}{TP+FN})$, where TP are the true positives and FP are the false positives.

Upload your solution on Open Data Playground:

Some specifications for the format of the solution that you will have to deliver on Open Data Playground. For each day of the challenge you will be asked to upload a .zip file to the platform with the following files inside:

(i) prediction .csv file (with a single column composed of the predicted values and header named “label”)
(ii) the reproducible code used to solve the problem
(iii) commented notebook to verify the reproducibility of the code

Furthermore, the .zip file must comply with the following formatting requirements:

(i) The zip file must contain the team name and cannot be used special characters and punctuation [i.e. teamname.zip]
(ii) The .zip file must not contain folders inside it otherwise it will not be accepted by the platform

On Open Data Playground you will have the possibility to upload different solutions to check which model is the most performing; the platform will only keep the highest score achieved. Additionally, after submitting a solution you must wait a period of 30 seconds before submitting a new one.

Overall hackathon rating:

Once the hackathon is concluded, a temporary ranking will be created resulting from the weighted average of the best results obtained by the team in the two days of the challenge.

Specifically, the best result obtained on DAY1 will count for 60% of the total score, while the best result obtained on DAY2 will count for the remaining 40%.