Data Science and Machine Learning Lab
Politecnico di Torino

Project Assignment
Summer Call, A.Y. 2025/2026

Last update: June 12, 2026

1 Project dates

Start date: June 09, 2026 at 23:59 (CET)
Due date: June 24, 2026 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Credit risk assessment is a fundamental task in the financial sector, where institutions must evaluate the likelihood that a client will fail to meet future payment obligations. In the context of credit card services, the ability to identify clients at risk of default is essential for reducing financial losses, improving credit limit management, and supporting responsible lending decisions.

The objective of this project is to develop and evaluate a machine learning model capable of predicting whether a credit card client will default on their payment in the following month. The prediction is based on demographic information, credit limit, historical repayment status, bill statement amounts, and previous payment amounts. The task is therefore formulated as a binary classification problem, where each client is assigned to either the default or non-default class.

2.1 Dataset

Warning: For this project, you are not allowed to use external datasets other than the one provided.

The full dataset contains 30,000 instances describing credit card clients. Each record corresponds to a single client and includes information about the client's credit limit, demographic characteristics, repayment history, bill statement amounts, and previous payment amounts over a six-month period, from April 2005 to September 2005. Each instance is associated with a binary target label indicating whether the client defaulted on the payment in the following month.

Several attributes characterize each record. The following is a brief description of each of them:

- ID: Unique identifier of the client.
- LIMIT_BAL: Amount of the given credit in New Taiwan dollars, including both individual consumer credit and family or supplementary credit.
- SEX: Gender of the client, where 1 indicates male and 2 indicates female.
- EDUCATION: Education level of the client, where 1 indicates graduate school, 2 indicates university, 3 indicates high school, and 4 indicates other education levels.

- MARRIAGE: Marital status of the client, where 1 indicates married, 2 indicates single, and 3 indicates other marital status.
- AGE: Age of the client in years.
- PAY_0: Repayment status in September 2005.
- PAY_2: Repayment status in August 2005.
- PAY_3: Repayment status in July 2005.
- PAY_4: Repayment status in June 2005.
- PAY_5: Repayment status in May 2005.
- PAY_6: Repayment status in April 2005.
- BILL_AMT1: Amount of the bill statement in September 2005.
- BILL_AMT2: Amount of the bill statement in August 2005.
- BILL_AMT3: Amount of the bill statement in July 2005.
- BILL_AMT4: Amount of the bill statement in June 2005.
- BILL_AMT5: Amount of the bill statement in May 2005.
- BILL_AMT6: Amount of the bill statement in April 2005.
- PAY_AMT1: Amount of previous payment in September 2005.
- PAY_AMT2: Amount of previous payment in August 2005.
- PAY_AMT3: Amount of previous payment in July 2005.
- PAY_AMT4: Amount of previous payment in June 2005.
- PAY_AMT5: Amount of previous payment in May 2005.
- PAY_AMT6: Amount of previous payment in April 2005.
- default.payment.next.month: Target label indicating whether the client defaulted on the payment in the following month.

The repayment status attributes PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, and PAY_6 describe the payment delay status for each month. Their values can be interpreted as follows:

- -1: Payment made duly.
- 1: Payment delayed by one month.
- 2: Payment delayed by two months.
- 3: Payment delayed by three months.
- 4: Payment delayed by four months.
- 5: Payment delayed by five months.
- 6: Payment delayed by six months.
- 7: Payment delayed by seven months.
- 8: Payment delayed by eight months.
- 9: Payment delayed by nine months or more.

The target attribute corresponds to the client's default status in the following month. The mapping between class names and their numerical labels is defined as follows:

- No default: 0
- Default: 1

The main challenge of the project is to build a reliable predictive model that distinguishes between clients likely to default and those not. Since the dataset is imbalanced, with the non-default class being more frequent than the default class, the evaluation should not rely only on accuracy. Metrics such as precision, recall, F1-score, ROC-AUC, and confusion matrix analysis should also be considered.

The dataset is located at:
`https://drive.google.com/file/d/1N07AVjwrKL2fNJNBfSlDef2ynhR7nlTq/view?usp=sharing`

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the records from the development set. This portion does have the `label` column, which you should use to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the records corresponding to the evaluation set. This portion does not have the `label` column.
- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a classification pipeline to predict whether each client in the Evaluation Set will default on their credit card payment in the following month.

2.3 Evaluation metric

Your submissions will be evaluated through Macro F1.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
0,0
1,1
2,1
3,0
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `id` in the evaluation CSV file.
- the `Predicted` label for the corresponding record.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_summer_2026**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in this form by the due date of this project. Failure to do so will result in a void project.

Warning: This is an additional requirement that was not required in past years. Make sure you remember to fill in the form by the due date, or your project will not be considered valid!