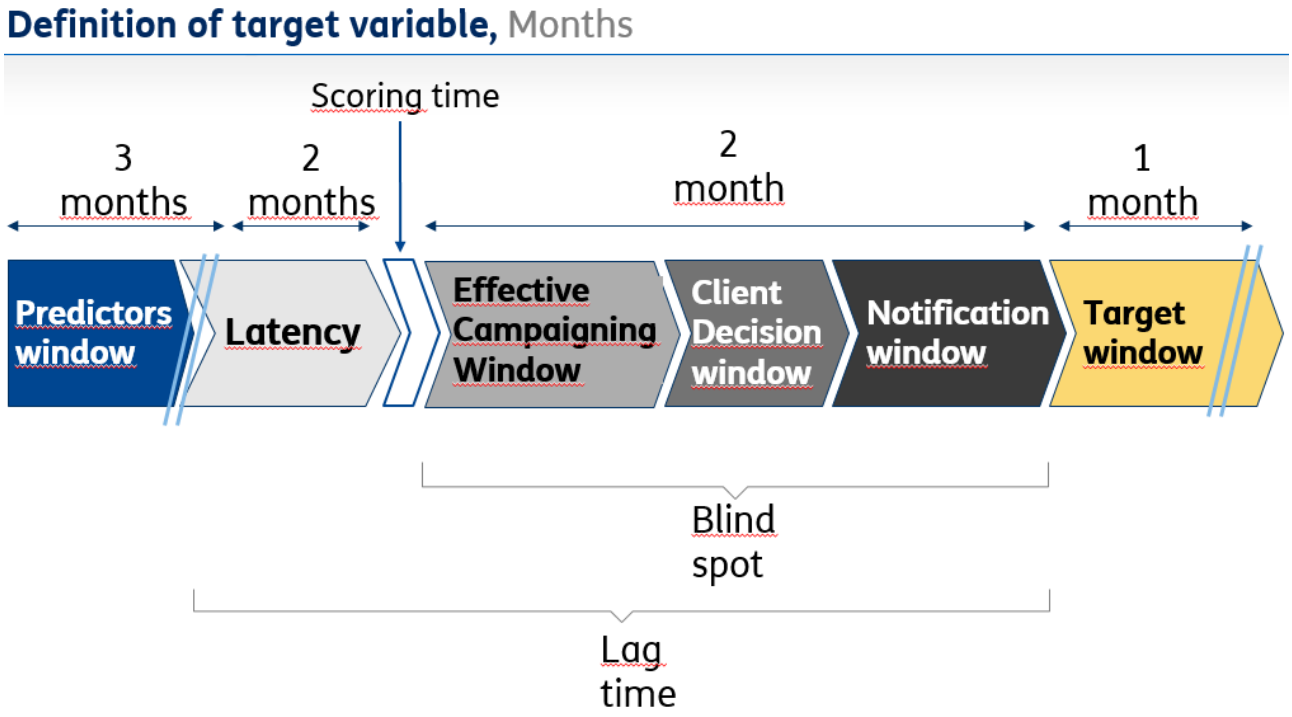
Instructions for participants

The data available are TIM's internal historical data, related to different domains (customer records, commercials, contacts, traffic, network quality, customer needs). The data time span is around 9 months. The data model, based on a realistic context, has the following properties:

- Dataset has a blind period (latency + blind spot) varying from 67 to 120 days depending on the variables (more details will be provided in *dataset_description.pdf*).
- *T_RUN* is the date which gives you the information about the timeslot of the data.
- Variables, whose name ends with the suffix *_DAYS*, contain measures up to the date of *T_RUN* - 7 days. Example: *CONT_IB_00_60_DAYS* indicates the inbound contacts that occurred in the last 60 days, starting from the date *T_RUN* - 7 days.
- Variables without suffix *_DAYS* are monthly variables.
- S_VAR specifies the reference month for monthly variables.
- The variable *S_TARGET* indicates the observed target month (*S_VAR* + 4 months or equivalent *T_RUN* + 2 months).

To understand the reference context, look at the data model shown below.



Latency is the delay affecting the received data.

Blind spot is the time window needed for the implementation of a retention campaign.

How to enter the competition

Each registered team leader will receive a link to access the material for the competition within 48 hours from the registration. Such link will expire at 23:59 of 01/10/2019.

The link will give access to the folder containing the following files:

- *dataset.csv* containing training data;

- *test_dataset.csv* containing the features/predictors but not the binary response variable, that will be retained by TIM for evaluation.

- *dataset_description.pdf* with all the information concerning the features and the metadata;

Terms of use: all the competition material is property of TIM and protected by Copyright: any use of the material other than the competition itself is prohibited.

How to submit

Registered teams will receive a link to upload, before 2/10/2019, the following files:

- *report_<teamname>*.pdf should contain:

    - Description of the problem and of the data, motivating the choice for external features from open data sources;

    - Choice of the procedure; in case more procedures are used, provide a possible explanation on why one procedure outperformed the others;

    - Measure of the features discriminant power;

    - Data visualization and unsupervised analysis as a support to the interpretation of the results.

- *prediction_<teamname>*.csv

    - the churn prediction as the probability to churn conditional to the observed values of the features contained in the file *test_dataset*.csv. The CSV file should adhere to the following format:

        COD_NUM_TEL,TARGET_PROB
        00327c214[…]803f81d3a2a,0.87
        00c7cb0d2[…]db4252f83a,0.23
        […],[…]

- *code_<teamname>.zip*

    R or Python scripts to reproduce the obtained results.

Note: replace <teamname> within the file names with the actual name of the team.