**Evaluation criteria**

The evaluation criteria for this competition include both technical and business aspects. Technical criteria are meant to assess the general ability of the teams to handle, interpret, and understand the data, build predictive models for the given problem and evaluate their performance according to an objective metric.

Business criteria are more general in scope, and they involve the ability to devise what value can be extracted from this data, formulate relevant business questions and try to find answers to them in the data. The quality of the presentation and the ability to communicate clear and powerful ideas on the final pitch will also be part of the business criteria.

Teams are required to submit:

1. the code they developed to tackle the problem (ideally with comments, or jupyter notebooks alternating code and text explanations).
2. An executive summary (3-5 pages) with their inspection strategy and value propositions.

The technical criteria will be assessed on the basis of the first submission, and the metrics obtained by the best model submitted to the Kaggle platform.  
Business criteria will be assessed mostly from the executive summary.

**Technical criteria**

Technical criteria will be evaluated on the code that each team is required to hand in for the first part of the competition. The code should be clearly structured and results should be interlaced with explanations in jupyter notebooks. The notebooks should be clearly written, and explain the process followed starting from the raw dataset, cleaning and preprocessing, exploratory data analysis, model formulation, hyperparameter tuning (if needed), final metrics and discussion.

Within the technical criteria, attention will be paid to:

* Overall understanding of the problem
* Exploratory data analysis
* Feature engineering
* Predictive models performance

**Overall understanding**

Ensure that you understand the meaning of each predictor variable: what it means, in which units is it expressed, how is this data registered, at which moment in the timeline of a given flight, could it contain errors? could it contain outliers ? can we trust the data ? Using common sense, will a given predictor variable be useful to predict our target?

**Exploratory Data Analysis**  
Getting acquainted with a dataset is a first necessary step before any modeling on the data takes place. Explore the data distribution, which variables are categorical and which are numerical, do we really understand the meaning of each variable? Are there any correlations among the variables ? Are there predictor variables with missing values or outliers ? Can we trust the values of the data ? Try to formulate hypotheses and understand your dataset before further exploration is conducted. Create good visualizations that help develop your intuition and understand the patterns. If necessary, decide how to handle missing values by either data imputation or removing rows/columns from the dataset.

**Feature engineering**  
Which features will you use in your predictive model ? Is it legitimate to use all the provided data? Can you imagine how the model will be used in production?

Can you enrich your dataset with external information ? At the very least you will need to merge the two datasets that you have and/or derive features to train models. Be creative: anything that you can build on the given data that might have a more direct connection to what you are trying to predict will improve your models performance.

**Model performance**

This will take the form of a Kaggle type competition. The proposed problem is a classification problem to estimate the risk of leakage in pipes following a scheduled inspection. Teams are required to submit their predictions into the Kaggle competition (maximum 10 submissions per day), and the models will be ranked with respect to the metric chosen for this competition, which is Area under the ROC Curve. Feel free to try different families of models, adjust their parameters, add regularization, go back to your preprocessing cycle and continue iterating, etc. You will use your training data to gauge the performance of your model and you will receive immediate feedback on the test set when you submit your predictions.

**Business criteria**

**Extracting value from data**

For this task you should assume that you have a prediction model that works reasonably well to predict the risk of failure in a pipe. Your task is how to use this model in a business case and how this brings value to the gas company.

With your model in hand, you should devise an inspection strategy, i.e. in which order should the pipes be inspected at every new year. For a realistic strategy, pipes cannot be visited in an arbitrary sequence over the whole territory, but subject to geographical constraints. The cost of pipe inspection is per unit length, and pipes have different lengths. You could assume that once a leakage is found, it has been leaking gas for half of the time since the last inspection (in the absence of further knowledge). The volume of gas leakage will also need to be estimated.

With all your estimated parameters in hand, evaluate the gain in efficiency produced by your strategy with respect to a standard rule based strategy that does not take into account predictive maintenance. You should be able to estimate a reduction in carbon footprint and also an economic gain for the gas company.

For instance, you could run your strategy over the years for which you have inspection data (2010-2020) and estimate what should have been a better sequence of inspections, as determined by your trained model (there’s a bit of overfitting in this argument, but that’s ok) and what are the benefits of your proposed strategy.

The main things we will value here is your ability to go beyond technical manipulation of data and model fitting in order to think from a higher perspective and understand how predictive maintenance can be used in practice, and whether its effects can be quantitatively estimated for decision making.

All 30 participating teams will be assessed on a combination of these two factors. A team with a model with poor performance might still qualify for the final if their business model and strategy is outstanding, so try to devote some time to both tasks. And remember to work in parallel and divide your team according to expertise and capacity.

**Pitch clarity**

The five teams that are selected for the final phase will pitch their results in front of a jury. Their technical results will have already been assessed by the technical jury on the basis of their submitted notebooks and model performance. On the pitch you will need to transmit clear and powerful ideas that highlight your results and show your understanding of the problem, your ability to harness value from the data and your ideas to contribute to the problem under consideration. Focus on the large scale goals, while showing evidence that your technical skills are solid, but do not use your time to explain straightforward technical solutions, unless you think that they are really essential.