Multi-Class Email Classification Challenge

2EL1730 - Machine Learning

Kaggle Challenge
January 2021

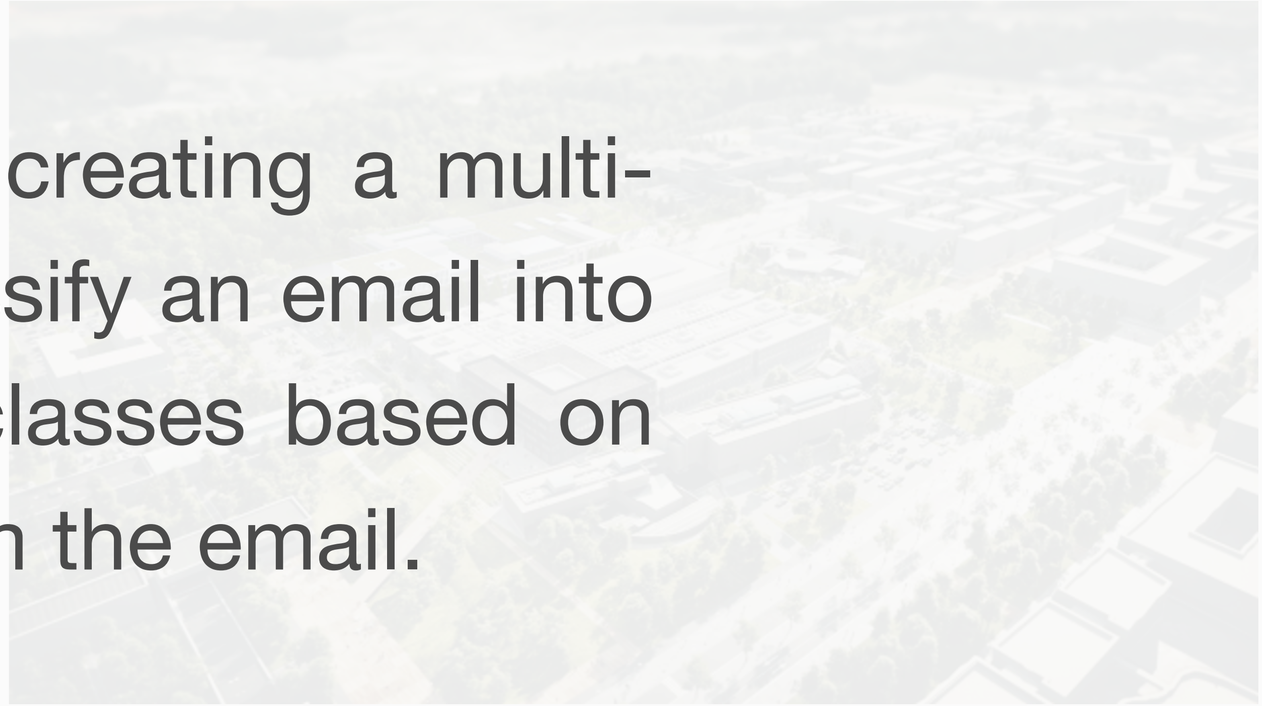
CentraleSupélec

Motivation

We often face the problem of searching meaningful emails among thousands of promotional emails.

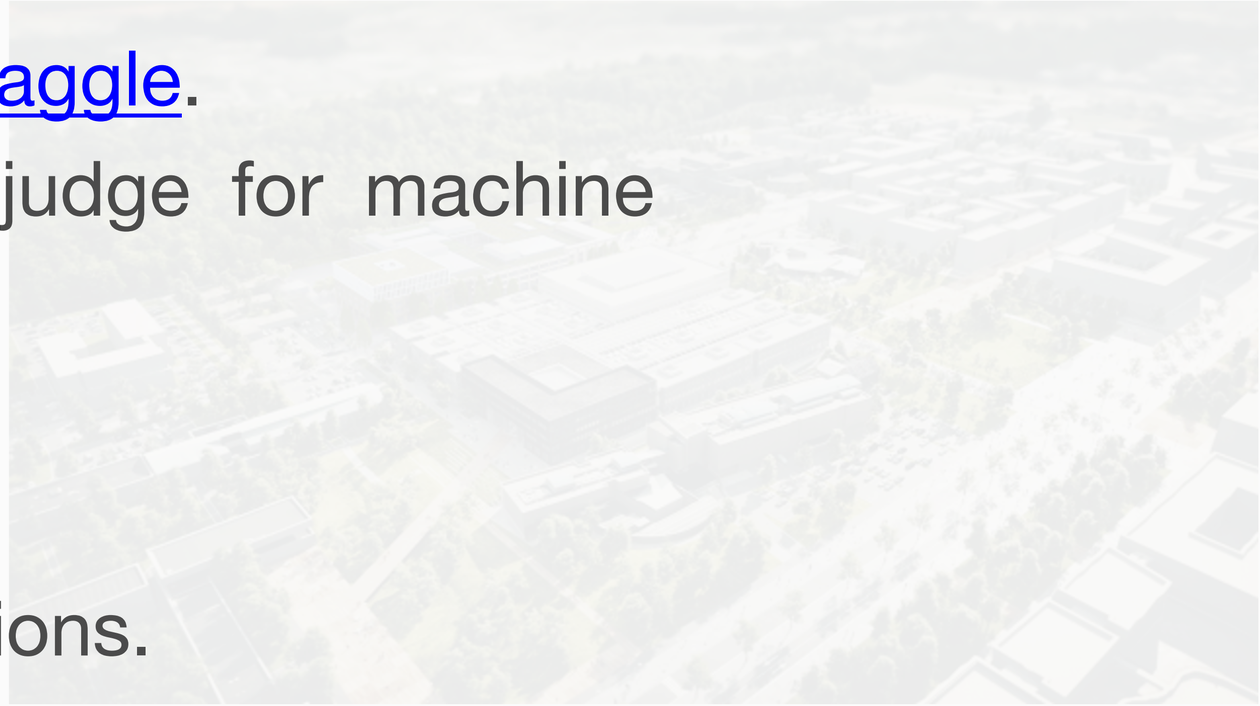
Challenge Goal

This challenge focuses on creating a multi-label classifier that can classify an email into one or more of the eight classes based on the metadata extracted from the email.



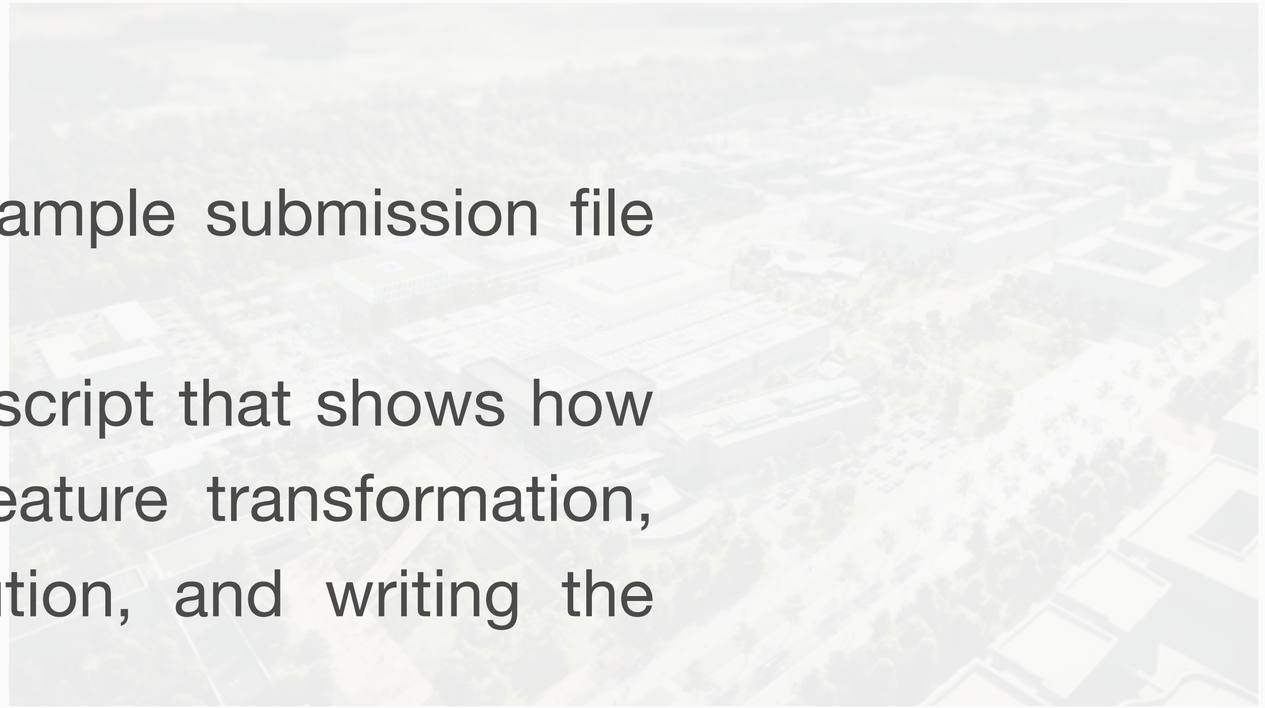
How to start with the challenge?

- The challenge is hosted on kaggle.
- Kaggle provides an online judge for machine learning problems.
- Register on kaggle.
- Go to the challenge.
- Accept the terms and conditions.



Files

- train_ml.csv - the training set
- test_ml.csv - the test set
- sample_submission_ml.csv - a sample submission file showing the correct format.
- skeleton_code_ml.py - a python script that shows how to read the data, how to do feature transformation, training a benchmark SVC solution, and writing the results to the submission csv file.



Dataset Features

- **date** - unix style date format, date-time on which the email was received, ***e.g. Sat, 2 Jul 2016 11:02:58 +0530***
- **org** - organisation of the sender, ***e.g. centralesupelec, facebook, and google.***
- **tld** - top level domain of the organisation, ***eg. com, ac.in, fr, and org.***
- **ccs** - number of emails cced with this email, ***e.g. 0, 2, and 10.***
- **bcced** - is the receiver bcc'd in the email. Can take two values 0 or 1.

Dataset Features (Cont.)

- **mail_type** - type of the mail body, ***e.g. text/plain and text/html.***
- **images** - number of images in the mail body, ***e.g. 0, 1, and 100.***
- **urls** - number of urls in the mail body, ***e.g. 0, 1, and 50.***
- **salutations** - is salutation used in the email? Either 0 or 1.
- **designation** - is designation of the sender mentioned in the email. Either 0 or 1.

Dataset Features (Cont.)

- **chars_in_subject** - number of characters in the mail subject, ***e.g. 0, 1, and 10.***
- **chars_in_body** - number of characters in the mail body, ***e.g. 10 and 10000.***
- **labels** - last eight columns represent eight classes, 0 means that label is not present for this row and 1 means that label is present, multiple label columsn can be 1. Label columns are only present in train.csv. test.csv has features only.

Class Labels (8 Classes)

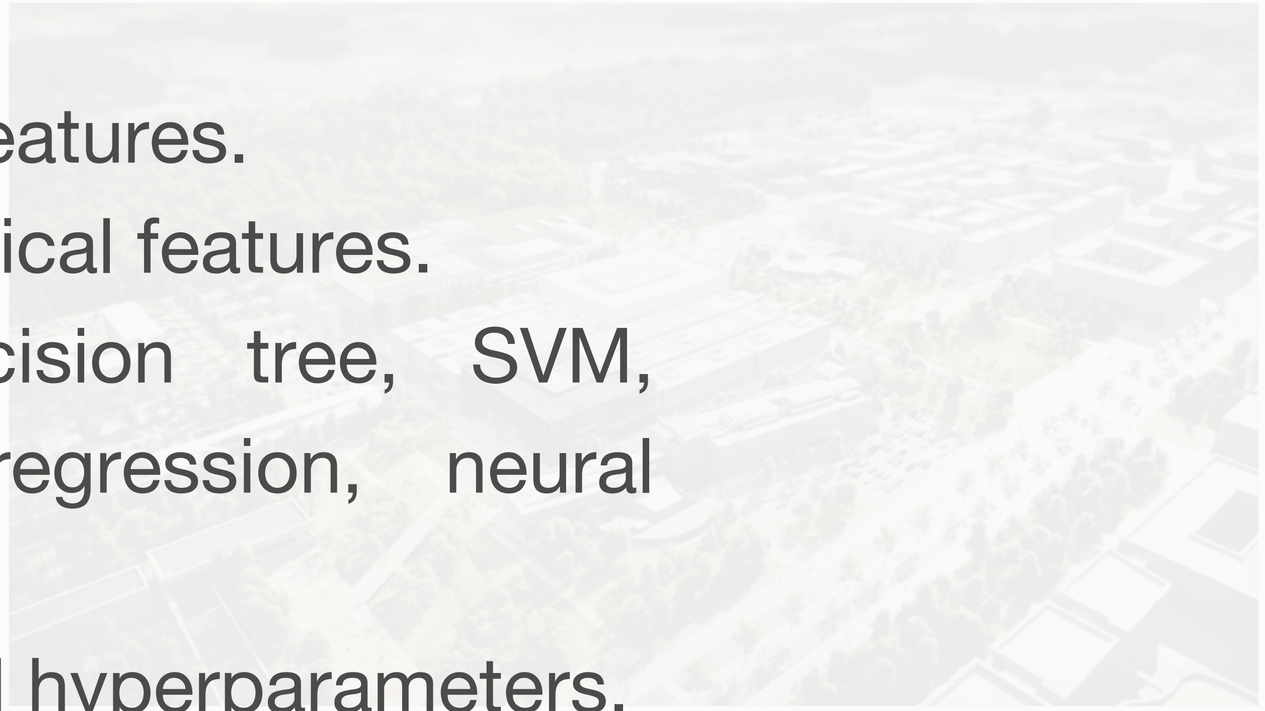
- updates
- personal
- promotions
- forums
- purchases
- travel
- spam
- social

Baseline Model

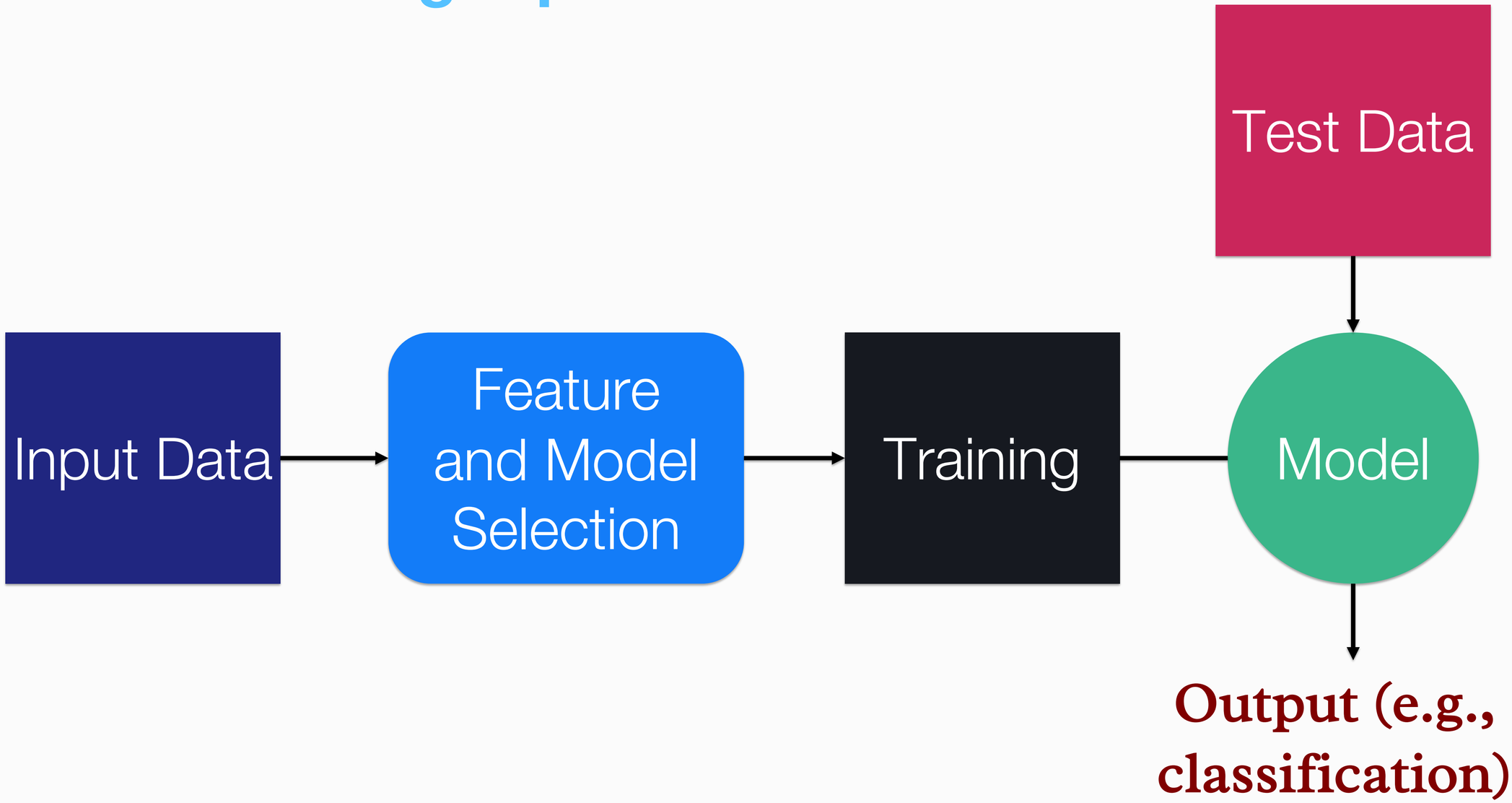
- SVC is used as baseline which performs worse than just predicting everything as 0.5.
- Only one of the feature ‘mail_type’ is used in the baseline.
- Mean Log-Loss for SVC is 0.30 and for sample submission it is 0.69.

Improving Baseline Model

- SVC with multiple features.
- Normalisation of numerical features.
- One hot encoding of categorical features.
- Trying other models: decision tree, SVM, random forest, logistic regression, neural network, etc.
- Grid search over models and hyperparameters.

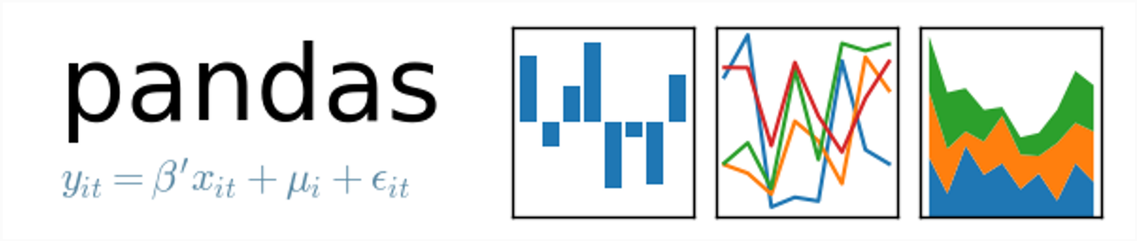


Machine Learning Pipeline



Software Tools

- Python libraries
- numpy
- scipy
- scikit-learn
- pandas
- anaconda includes almost all the required packages

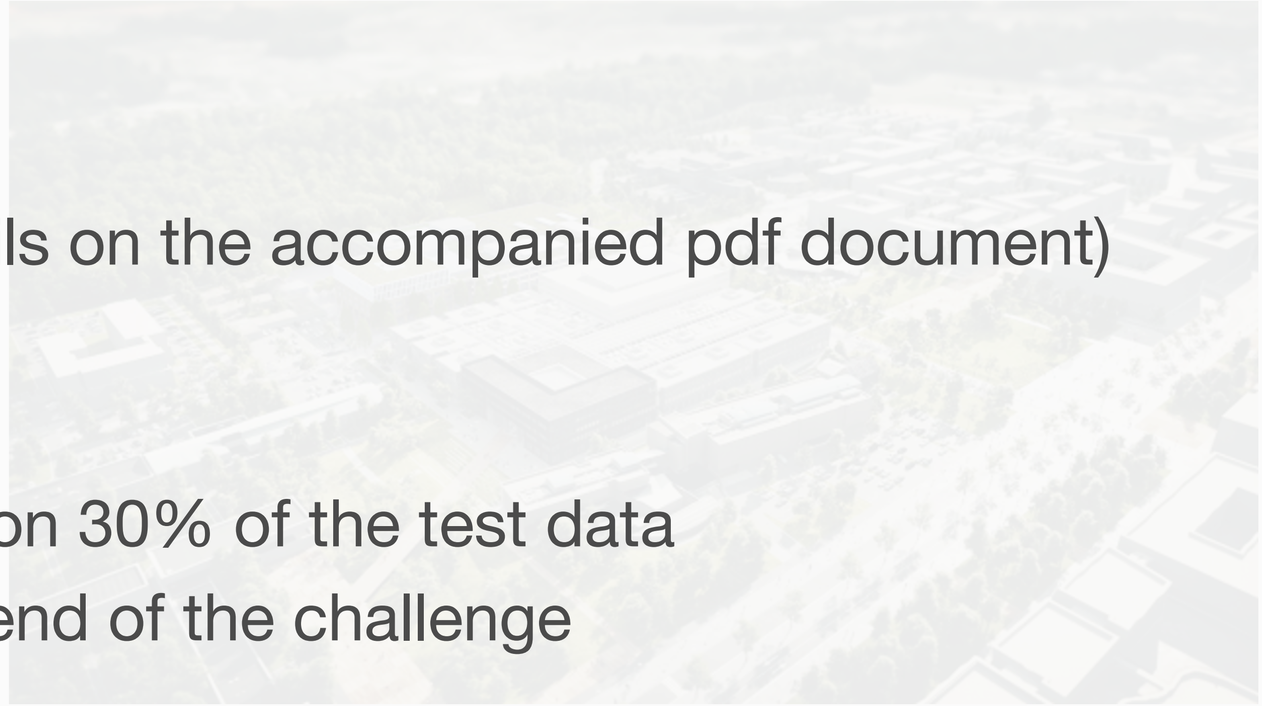


Submission Details

- Submission on kaggle (see the details on the accompanied pdf document)
- Your best performing model
- Leaderboard score

  Public: what you see - computed on 30% of the test data

  Private: will be announced at the end of the challenge



Deadline: January 31, 2021

- 06:00 PM: Submission deadline
- For any help contact Sagar

  Email: sagar.verma@centralesupelec.fr

Good Luck and Enjoy!