Project for Machine Learning I

UMONS 2023-2024
Victor Dheur (victor.dheur@umons.ac.be)
Tanguy Bosser (tanguy.bosser@umons.ac.be)

1 Overview

The objective of this project is to train the most accurate regression model using a dataset composed of 1459 observations with 35 continuous predictors and 43 categorical predictors. Each observation represents multiple characteristics of a house, and your task is to predict the lot area of the house. The dataset originates from the following Kaggle contest [1]. Participants are prohibited from using any information outside this dataset, or any other dataset, for model development. Teams must train at least two regression models: one will be directly assigned to you, while you must independently choose the other(s). After the model development phase, teams are expected to submit their models for the competition, detail their methodologies and results in a comprehensive report, and articulate their findings in an oral presentation.

The initial dataset has been split into training and test sets. The full training set is available, but only the predictors are provided for the test set. You can evaluate your predictions by submitting them to the Kaggle website. Only a random subset of 50% of the test set is used to compute your *public score*. Your final *private score* using the remaining 50% of the test set will be provided at the end of the competition.

The evaluation metric for this competition is the root mean squared error between the logarithm of the true value and the predicted value:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log \hat{y}_i)^2},$$

where:

- n is the number of data points in the test set,
- y_i is the true value of the lot area,
- $\hat{y}_i$ is the predicted value of the lot area.

2 Kaggle competition

The competition (with related datasets) is accessible through the following invitation: `https://www.kaggle.com/t/7dcba15eea084125820b8db77cdd65ba`.

The following tasks show how to participate in the competition:

1. Form a team composed of three people.
2. Each team member should create a Kaggle account (using his/her UMONS email address).
3. Form a team on Kaggle.
4. Do some basic exploration of the dataset.

[1] `https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques`

5. Build your first model and upload your predictions to Kaggle. Your predictions must follow a certain format and must be contained in a .csv file (see `https://www.kaggle.com/code/vekteur/sample-submission`).

6. Try, and try again to improve your model. You can make a maximum of five submissions per day.

   To help you, a "get started" notebook is available at `https://www.kaggle.com/code/vekteur/sample-submission`. However, do not follow it blindly, and make sure that your implementations follow the methodology seen in class!

3 Report

The report (ideally written in LaTeX) can be a maximum of **10 pages** in **PDF format** and can be written in either French or English. If you write it in French, do not try to translate all the technical terms (e.g., cross-validation should preferably remain cross-validation). The report must abide by the section structure described below.

1 **Section 1: Exploratory Data Analysis** (max 2 pages). In this section, you must investigate the data to detect patterns and spot anomalies that might be present amongst the different variables, through graphical representations (scatter plots, boxplots, etc...).

2 **Section 2: Methodology**. This section describes the models/methods you have used, including a justification of your choices. You should also present your model fitting, diagnostics, etc. You should discuss and compare the model assigned to your team and at least one other model.

3 **Section 3: Results and Discussion**. In this section, you must discuss your classifiers' results, through the means of tables, figures, etc... You should clearly explain the performance you obtained for your models.

Overall, you will be graded based on clarity of writing, quality of presentation, level of machine learning content, and technical communication of main ideas. You should clearly explain what you have done, using figures to supplement your explanation. Your figures must be of **proper size with labeled, readable axes**. In general, you should take pride in making your report readable and clear.

4 Code

Your code should be properly structured and well-commented, and your experiments should be **reproducible**. You may choose to submit your files as .py or .ipynb files.

5 Oral presentation

Each team will present their project during a **10-minutes** presentation, followed by questions from the teaching staff. The presentation should summarize the main experiments and results following your project report. It should include the model assigned to your team and at least one other model. Each member of the team should talk for about 3 minutes. Keep in mind that your presentation will be **stopped** after 10 minutes. Your slides should be in **PDF format**.

6 Grading

- Total points: 20
- Accuracy of your final classifier on Kaggle: 6
- Report, code, and presentation: 14

7 Deadlines

- **April 14, 11:59pm**: Your team should be formed on Moodle.
- **April 28, 11:59pm**: At least one Kaggle submission needs to have been made.
- **May 12, 11:59pm**: The Kaggle competition closes.
- **May 15, 11:59pm**: Upload to Moodle a single zip file containing your project **report** (PDF), **code** (.py/.pynb files) and **presentation slides** (PDF), one per group.
- **May 17, 9:30am-12:30pm**: Oral presentations. Individual team schedules will follow.

Do not wait till the last minute. Late submissions will not be allowed.