UNIVERSITÉ LIBRE DE BRUXELLES ULB

FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

INFO-F-422 - Statistical foundations of machine learning
Project 2022-2023

Gian Marco Paldino, Cédric Simar, Gianluca Bontempi

April 27, 2023

The project counts for 50% of your final grade (i.e. 10/20). This project has to be developed by a team of 2 students registered for the class. Any project returned by a team composed of a different number of students will not be considered. The project shall be completed independently and it shall represent the sole efforts of the team submitting the assignment. The result of another team's efforts, or the copy of another team's efforts (current, or past, semester(s)), is considered academic dishonesty and will be punished accordingly. The assistants are aware of the existence of publicly available Jupyter notebooks for the competition.

1 Goal

The goals of the project are:

- To participate to the *"Richter's Predictor: Modeling Earthquake Damage"* DrivenData competition by implementing and assessing different supervised learning algorithms and different methods of feature selection in the related classification task.
- To select among the learning and feature selection techniques the ones that are considered to generalize better on the test set, with a thorough discussion.
- To report analyses and results as a Jupyter notebook, commenting on the scores obtained in the DrivenData competition and their relationship with the previous discussion.

2 DrivenData competition

The goal of the competition is to predict the level of damage to buildings caused by the 2015 Gorkha earthquake in Nepal, based on aspects of building location and construction. The data was collected through surveys by Kathmandu Living Labs and the Central Bureau of Statistics, which works under the National Planning Commission Secretariat of Nepal. This survey is one of the **largest post-disaster datasets ever collected**, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics. The dataset includes roughly 260.000 labeled samples and 38 features.

The model has to be trained using the *Train values*, *Train labels* files available on the DrivenData platform (see Figure 1). The students should then predict the *labels* for the samples included in the *Test values*, and submit them to the platform following the provided *Submission format*. The students should register **using its ULB netid as username** and accept the rules of the competition (notably no hidden additional accounts).

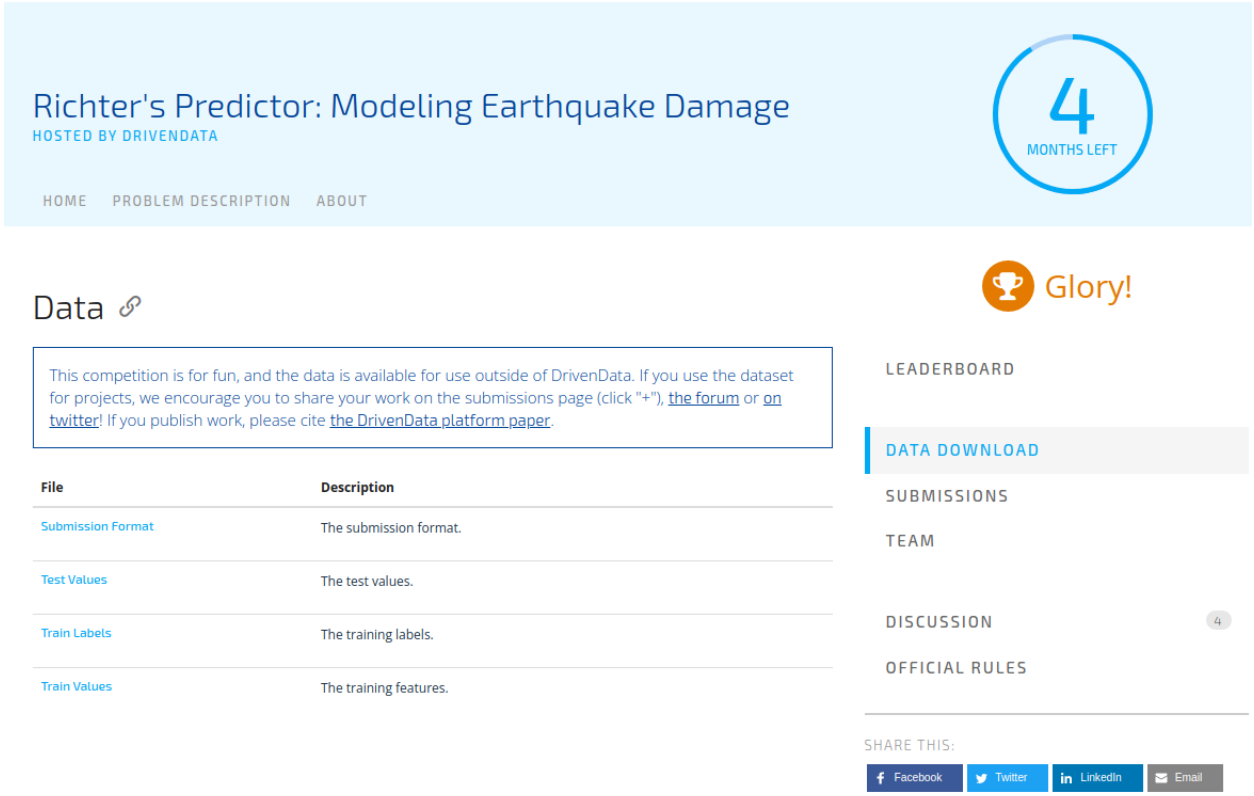

Figure 1 – Screenshot of the Data Download section of the DrivenData platform

3 The team

A project team has to be composed by exactly 2 ULB students registered to the class. Projects submitted by teams composed by a different number of students or by a student not officially registered will not be considered. Team composition must be submitted **by one of the group members** through a dedicated Google Form, available **here**. The team composition has to be finalized no later than **11PM of May the 1st 2023**. (Hard deadline!)

4 Tasks

The team will have to:

1. implement in the R language a pipeline for data preprocessing, including missing value imputation, normalization (if required), feature engineering and feature selection. This procedure must be detailed in the notebook. The text must contain the list of relevant/selected variables and the motivation of their choice. The use of visualizations and tables to provide a better understanding of the data and the usage of formulas and pseudo-code to describe the feature selection procedure is strongly encouraged. Note that one-third of the score will be assigned according to the quality of the documentation. **(2 points)**

2. implement in the R language a model selection procedure to compare at least three strategies for dealing with the multi-class nature of the prediction problem. This procedure must be detailed in the notebook and **exclusively use the packages listed in Section 5**. The text must mention the different (and at least three) strategies (among those presented during the course) which have been taken into consideration and the procedure used for assessment and selection. The use of figures, formulas, tables and pseudo-code to describe the model selection procedure is strongly encouraged. Note that one-third of the score will be assigned according to the quality of the documentation. **(2 points)**

3. implement a learning procedure **using other R packages than the ones listed in Section 5** . This procedure must differ from the one in the previous point in terms of the classification model (e.g. a deep learning model, a gradient boosting tree) and/or the feature selection strategy. A procedure combining multiple models is also allowed, provided that it integrates at least a learner different from the ones presented in the practicals. This procedure must be detailed in the notebook. The text should justify the choice of this procedure, assess its accuracy with respect to the one developed in point 2 and discuss the results. The use of figures, formulas, tables and pseudo-code to describe the combination of this novel procedure is strongly encouraged. Note that one-third of the score will be assigned according to the quality of the documentation. **(2 points)**

4. using the classifiers learned in the previous steps, the team must compute the predictions for the competition and submit them via the DrivenData website. The name of the team should appear on the official leaderboard of the competition. The link to enrol in the DrivenData competition is available **here**. **(1 point)**

5. using one of the classifiers learned in the previous steps, the team must return a ranking of the features in terms of relevance. The relevance of the ith feature must be computed as

$$\mathrm{H}(\mathbf{y}|\mathbf{X}^{-i}) - \mathrm{H}(\mathbf{y}|\mathbf{X})$$

where $\mathrm{H}(\mathbf{y}|\mathbf{X})$ is the conditional entropy of the target $\mathbf{y}$ given $\mathbf{X}$ and $\mathbf{X}^{-i}$ is the set of variables of $\mathbf{X}$ with $\mathbf{x}_i$ set aside. The students should refer to the definition of entropy of a discrete random variable in the Section 2.11.3 of the Handbook. **(3 points)**

5 Specifications

The team has to choose a learning method and a feature selection method among at least three alternatives. For the learning method (at point 2), the only packages that may be used are those included in this list :

- stats/ridge (linear/ridge models)
- keras (neural networks)
- tree/rpart (decision trees)
- randomForest (random forest)
- RSNNS (radial basis functions)
- lazy (nearest neighbours)
- e1071 (SVM)
- glmnet (LASSO/ElasticNet models)

For the point 3. the team is free to employ **other learning methods**, either already available online, or coded. The report must be an R Jupyter notebook which has to specify and justify (with tables, figures) the selection procedures which led to the final choice. The team has to return, together with the report, the datasets employed in the notebook, the set of predictions submitted to the DrivenData competition and a video summarizing the whole predictive procedure. **Important**: do not forget to set the seed of R's random number generator, it will ensure that we are able to re-execute your code and obtain the same results.

6 Performance metric

We are predicting the level of damage from 1 to 3. To measure the performance of our algorithms, we'll use the **F1 score** which balances the precision and recall of a classifier. Traditionally, the F1 score is used to evaluate performance on a binary classifier, but since we have three possible labels we will use a variant called the **micro averaged F1 score**.

$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \quad R_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3} (TP_k + FN_k)} \quad P_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3} (TP_k + FP_k)}$$

Students are not required to implement this performance metric manually. An implementation can be found in the following R package: **MLmetrics**. An example of its usage can be found here.

7 Github

The submission of the deliverables for the project is made on a group repository on the GitHub Classroom platform. **The access link required to setup your repository on the platform will be provided once all the teams are formed.** Once you click the link, you shall either create and name your group (i.e. "Group [group number]") or join your existing group if it has already been created by your group partner.

The version control system offered by Git allows you to easily keep track of the different versions of your code, and go back to a previous version of your code in case you would have any problem. In order to do so, you will need to regularly instruct the system to keep track of the different version of your code (i.e. `commit` in the Git terminology). However, a `commit` only keeps track of a version of your code locally (on your PC). In order to transfer a copy of your code to a remote server (like Github) you will need to perform a `push` operation. We encourage you to regularly perform `commit` and `push` operation in order to have up-to-date versions of your code both locally and remotely. More information concerning the Git workflow and how to manage your repositories can be found at `https://docs.github.com/en/repositories`. A short kick-start demonstration of Git and Github Classroom will also take place during the practical session introducing the project.

For each group of students, the final submission of your code that will be evaluated is the last uploaded version of the deliverables on your assigned Github repository before the deadline time and date. If you wish to submit your final deliverables between the 11PM of May the 19th 2023 deadline and the 11PM of May the 20th 2023 (penalized by one point), you need to contact your teaching assistants to make sure we evaluate the correct deliverables.

8 Students participation

The Git version control system will also be used to assess students participation in the group effort. Students with little to none visible participation in the project will be penalized accordingly. Thus, we encourage every member of each group to regularly perform `commit` and `push` operations so that we can reliably appreciate the input of each member with respect to the final deliverables.

9 Plagiarism and code generation

Considering that many notebooks have already been published about this competition, and that new code generation tools were recently made publicly available, we would like to especially draw your attention to the fact that this project **shall be completed as a team and it shall represent your sole efforts**. Each project will be scanned for text and code plagiarism or automatic generation. The result or the copy of another team efforts (current, or past, semester(s)), of a publicly available notebook, either in full or in part, or the use of an automatic code/text generation tool, is considered academic dishonesty. Plagiarism, in the sense of copy-pasting from existing, or generated, reports or code is a serious issue and will lead to disciplinary measures being applied on the whole group.

10 Deliverables

The student team will deliver its assigned Github repository:

1. the implementation (in a R Jupyter notebook format) of the preprocessing pipeline, model selection and predictive procedure.
2. the datasets employed in the notebook (in .csv format).
3. the predictions submitted to the competition (in .csv format).
4. a video of max 10 minutes to present this project. The presentation should address the main points illustrated in the Jupyter notebook. Each student has to present half of the project. Given the size of the videos and the format of the video, the video may need to be stored on a different platform (e.g. Microsoft Stream, Youtube). In this case, please provide a link to the video of the presentation. In case of problems, get in touch with an assistant (`gpaldino@ulb.ac.be`) to find an alternative solution.

A template describing the structure of the project is available on the UV in the Project section as well as inside the Github repository that will be automatically created for your group.

Rules for project submission

To be read carefully!

1. The assignment should be made by teams of **exactly** two students. The team composition has to be finalized no later than **11PM of May the 1st 2023**.
2. The assignment will be graded on the implementation, the report and the video presentation.
3. The code should be **commented**.
4. The assignment will be submitted through the Github Classroom platform and the repository should include all deliverables listed in the deliverables section.
5. The deadline for the submission of your project is: **11PM of May the 19th 2023**.
6. All the projects submitted after the deadline will be:
   - Penalized by one point if submitted **before 11PM of May the 20th 2023**.
   - Not considered **after 11PM of May the 20st 2023**
7. Sharing of code is not allowed (you may, however, verbally discuss ideas on how to tackle the project).
8. This project counts for 50% of your grade (10 points). This project **shall be completed as a team and it shall represent your sole efforts**. The result or the copy of another team efforts (current, or past, semester(s)), or automatic code generation, is considered academic dishonesty. Plagiarism, in the sense of copy-pasting from existing reports or code is a serious issue.
9. Each project producing any error during its execution will receive a grade of 0/10.