Guidelines

During the *Data Analysis for Business* course, we have introduced you to different tools intended to perform three different kinds of tasks:

- **regression:** prediction of a numerical (continuous) outcome;
- **binary classification:** prediction of a binary (discrete) outcome;
- **clustering:** partition of the observations into homogeneous groups.

In the development of the **final project**, we expect you to pick a dataset and complete a **primary task**, either *regression* or *classification*, denoted as **Task 1**. In case the dataset is suitable for an additional task, you are free to consider the implementation of a **Task 2** (secondary task, the remaining one or clustering). These tasks shall be completed having a meaningful objective in mind, i.e. drawing sensible conclusions in terms of marketing advice, policy making, etc.
In particular, the process must follow the steps here below.

1. **Pick a dataset**. Suitable datasets will be provided on the *LUISS Learn* webpage. Nevertheless, you are free to choose any other dataset and discuss with Prof. Iafrate and/or the TAs the actual feasibility of your choice.

2. **Describe the dataset** at your best. Provide a general overlook (size of the dataset, features characterization) and ***clean/deal with*** dirty records (e.g. duplicates, missing values, unary variables). Show some plots and descriptive statistics. This does **not** mean describing **everything**, but only those features and characteristics that you consider relevant to the general understanding of the data and to the end of the subsequent analyses.

3. Define the **objective** of **Task 1** (e.g. regression on variable Y) and explain how this can be relevant in practical terms. The provided datasets will come with an example objective: you are free to pick a different one if you wish.

4. Before proceeding to the full model (aimed at providing the best result), focus on some **lower-dimensional models** (considering only a few variables) in order to investigate some relationships you consider interesting (e.g. what we did with `Total_Trans_Amt` and `Total_Trans_Ct` in the *midterm*).

5. **Do your best.** Split the data in a training and a test-set. Taking into account all the techniques we introduced during the course provide your best solution (the best model in validation) for completing **Task 1**. You can also consider any *feature engineering* you like. In particular, we want you to explicitly compare at least one technique from each of the following classes: *linear models* (AIC or BIC step-selection), *penalized approaches* (Ridge, Lasso, Elastic Net), *non-linear models* (knn, splines, gam, tree-based algorithms).

6. **Draw your conclusions.** Evaluate the performance of your model on the test-set using the various metrics we introduced and draw some conclusions on the structure of the chosen model: what variables were selected? Is there any interesting linear or non-linear relationship?

7. **In case you decided to go for the additional task.**
   Define the **objective** of **Task 2** (e.g. classification on another variable, or clustering) and explain how this can be relevant in practical terms. Repeat steps 5 and 6 for **Task 2**.
   In case you are going for *clustering*, you shall compare k-means and hierarchical methods. Pick the best number of clusters using any of the tools we introduced (WSS, Silouhette, Gap-Statistic) and visualize the resulting clusterization in a lower dimensional space. If possible, compare your clusterization with possibly available labels that were not included in the clustering process (as we did for the products and the regions in the *midterm*).

Submissions

1. Write a report/notebook in whatever format you prefer, but the **highly recommended** option is to use a `RMarkdown` file in order to automatically include figures and code in the report. This **must not include** all the elaborations you did, but only the **most relevant ones** (those you consider worth explaining in detail). The irrelevant elaborations may be cited (e.g. *"we also compared the results obtained with AIC-forward/backward step selection, but these lead to worse validation scores"*), but without any need to including all the details. Please, submit both the .Rmd file and the compiled version. The compiled version (pdf or html) must **not exceed 7 pages** (if you are doing only one task) or **10 pages** (if you are doing 2 tasks).
2. Include the original `R` code (or codes) you used to perform **all the elaborations**. Try to extensively comment your code to make it more readable. These codes will be checked in case something you wrote on the report did not convince us (so... make the report as convincing as possible!)

The **final report** and **codes** must be submitted by the group leader using the dedicated form on the LUISS learn course page.

Grading

Be aware that, when evaluating the performance of your final model, we do not really care how well the model performs on the test set. This is highly data dependent and out of your control. The final grade will be based on:

- **statement of the objective and choice of the appropriate task**;
- **correct use and interpretation of the considered techniques**;
- **quality of the final report**;
- **proficiency and ability to answer possible questions during the presentation**.

As usual, higher grades will be granted for insightful comments, smart coding and clever data visualization.