Data Science and Machine Learning Lab
Lab 10 - Regression

Politecnico di Torino

Intro

The main objective of this laboratory is to put into practice what you have learned on regression techniques. You will work on a tabular dataset. In particular, you will build a regression model to predict the price of an Airbnb apartment based on various listing details.

Important note. As far as this laboratory is concerned, you are encouraged to upload your results to the competition we launched on our platform, even if the submission will not count towards your final exam mark. You have to use the same personal key you already used for Lab 9. If you do not have a key yet, please write to lorenzo.vaiani@polito.it. Refer to Section 3 to read more about the competition.

Important dates

Start date: December 15, 2025 at 10:00 AM (CET)
Due date: December 23, 2025 at 00:00 AM (CET)

1 Preliminary steps

1.1 Datasets

In this laboratory, you will use a publicly available dataset. Public Domain Dedication datasets constitute an extremely valuable asset for the data science community. If you want to know more about how they are distributed, refer to the CC0 licence.

1.1.1 New York City Airbnb Open Data

This public dataset is part of Airbnb, and the source can be found on Inside Airbnb.

Each row of the dataset corresponds to an Airbnb. As for the previous competition, the dataset has been divided into a Development set and an Evaluation set. You will find more about them later in the document.

- id: a unique identifier of the listing
- name: listing description
- host_id: a unique identifier of the host
- host_since: date when the host created the account
- host_location: location declared by the host
- host_response_time: typical response time of the host

- host_response_rate: percentage of messages the host responds to
- host_acceptance_rate: percentage of booking requests the host accepts
- host_is_superhost: whether the host is a superhost (boolean)
- host_total_listings_count: number of listings managed by the host
- host_has_profile_pic: whether the host has a profile picture (boolean)
- host_identity_verified: whether the host's identity is verified
- neighbourhood: neighborhood where the listing is located
- district: smaller administrative subdivision of the city
- city: city where the listing is located
- latitude: coordinate expressed as a floating point number
- longitude: coordinate expressed as a floating point number
- property_type: type of the property listed (e.g., apartment, house)
- room_type: type of room available to guests
- accommodates: number of guests the listing can host
- bedrooms: number of bedrooms available
- amenities: list of amenities offered by the listing
- minimum_nights: minimum nights requested by the host
- maximum_nights: maximum number of nights the guest can stay
- review_scores_rating: aggregated rating of the listing (available also separately)
- instant_bookable: whether the listing can be instantly booked (boolean)
- price: price per night expressed in dollars

You can download the dataset at:

`https://dbdmg.polito.it/dbdmg_web/wp-content/uploads/2025/12/Airbnb.zip`

1.1.2 Dataset tree hierarchy

The data have been distributed uniformly in two separate collections. Each collection is in a different file. The dataset archive is organized as follows:

- `development.csv` (Development set): a collection of listings **with** the price column. This collection of data has to be used during the development of the regression model.
- `evaluation.csv` (Evaluation set): a collection of listings **without** the price column. This data collection has to be used to produce the submission file.
- `sample_submission.csv`: a sample submission file.

So far, you should be used to working with training, validation, and test sets while developing your models. In this case, the Development data should be used to tune your hyperparameters, and you should treat the Evaluation portion as the actual test set.

2 Exercises

In this laboratory, you have a single regression task to carry out.

2.1 Airbnb listing price regression

In this exercise, you will try to predict the price of an Airbnb listing using several contextual information. To do so, your primary goal will be to model, using a regression-based pipeline, the relationship between listing information (e.g., its geographical location, reviews, or other metrics you identify) and the listing price.
Once your model is complete, you will predict, for a set of listings whose prices are unknown, how much it would cost you to spend one night at each.

Finally, you will be able to upload your regression results and participate in the lab competition.

1. Load the dataset from the root folder.
2. Focus now on the data preparation step. You should have noticed that the attributes that describe each listing are heterogeneous, both in their sources (e.g., geographical, related to the host, related to Airbnb, etc.) and in their types (e.g., numerical, categorical, date, etc.). Before continuing, take your time to answer these questions:
   - which attribute (or set of attributes) you think could drive the price per night the most?
   - can you detect any irregularity in any attribute distribution?
   - if your regression model will fit on numerical data only, how could you handle categorical attributes or text?

   Transform your initial dataset following the ideas you draw out.
3. Once you have your final dataset representation, choose one regression model from those you know. Then, use the classic training-validation pipeline on the Development dataset to identify the best hyperparameter settings for your model. As you can read in Section 3.3, we will evaluate your results on the MSE score (Mean Squared Error). Hence, it is a reasonable option to try to optimize it on the Development set.
4. Assign a price value to each listing in the Evaluation set.
5. Define a function to generate a 2D scatterplot with the prices. The chart must be drawn as a heatmap: use the latitude and longitude coordinates along the axes and the price value to assign a color to the point. Then, apply the function to the prices from the Development set and to the ones you predicted for the Evaluation set. From Section 1.1, you know that Development and Evaluation were generated with a uniform sampling on the initial listings. So, what should you expect on the map if your regression were correct?
6. Upload your results to the submission platform. Head to Section 3 to know more about it.
7. Compile your final report (if you want feedback) and upload it to the "Portale della Didattica" as described in section 3.2.

3 Submitting you work

For this laboratory, you should upload two files to two different websites. The first file contains the regression results, the second file contains a report on the experiments you carried out. The following sections provide further details on that.

3.1 Submit your classification results

To get your results evaluated, you have to upload a result file on our submission competition. The submission file has to be a `.csv` file formatted as follow:

```
Id,Predicted
10,120.31
123,100.00
21,523.22
345,652.02
42,225.41
...
```

As you can see, it must contain a header line and a row for each listing in the Evaluation collection. Each row must have two fields:

- the `Id` of the listing, as an integer number. Note that IDs can be in any order, but they **must** match the IDs present in the Evaluation set (column `id`).
- the `Predicted` price value, as either a float or an integer number.

The submission platform is the same you used for Lab 9. Therefore, you have to use the same key. Please refer to the guide on the course website, to go through the submission procedure.
You can find the competition at http://trinidad.polito.it:8888

3.2 Upload your report (optional)

For those interested, it is possible to submit a report describing the proposed solution for this laboratory. If you would like to receive feedback about your work, please contact lorenzo.vaiani@polito.it **before December 19**, explicitly indicating that your report has already been submitted and that you are interested in receiving comments on the report.

Please respect the following requirements:

- state clearly which pre-processing step characterized your final solution;
- describe which regression algorithm you used;
- describe which validation strategy you adopted and which are the best hyper-parameters you found on the Development set.
- comment on the heatmaps obtained in Point 5. Do they have the same "heat" distribution? If so, why? If not, why?

Please refer to the directions provided during the dedicated lecture to write your report. More specifically, you should use the IEEE conference LaTeX template. That is the template you will be using for submitting your final (graded) report, so you should get acquainted to it (and to LaTeX in general).

You can upload the file to the “Portale della Didattica”, under the Homework section of the course. Please use as description: `report_lab_10`.

3.3 Evaluation

Your regression results will be evaluated on the R2 score.