Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
Summer Call, A.Y. 2024/2025

Last update: June 5, 2025

1 Project dates

Start date: June 4, 2025 at 23:59 (CET)
Due date: June 19, 2025 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Accurately predicting apartment rent prices in urban housing markets is a crucial aspect of real estate analytics, with significant implications for property management, investment strategies, and housing policy planning. Reliable rent price forecasts can inform prospective tenants and landlords about fair market prices, help investors evaluate property profitability, and guide urban planners in addressing housing affordability challenges. By anticipating fluctuations and trends in rental markets, stakeholders can make informed decisions that align with their financial goals and social responsibilities.

The primary goal of this project is to develop a machine learning model that accurately predicts the monthly rent price of an apartment based on various features such as location, size, number of rooms, amenities, and neighborhood characteristics. This predictive capability is essential for creating data-driven insights into rental housing markets and ensuring transparency and fairness in rental pricing practices.

2.1 Dataset

The dataset used in this project comprises 99,487 apartment rent announcements collected from various real estate and classified listing sources across the United States. Each record represents a unique apartment. This dataset's blend of numerical, categorical, and textual data provides a comprehensive view of the rental housing landscape, enabling nuanced prediction of apartment rental prices that incorporate quantitative factors (like square footage and number of rooms) and qualitative factors (like apartment descriptions and amenities). Such richness is valuable for constructing a robust machine learning model that captures the diverse factors influencing rent prices across different neighborhoods and property types. The dataset includes the following features:

- **id**: Unique identifier for each apartment listing.
- **category**: The classification of the listing within the rental market.
- **title**: Short textual headline summarizing the apartment.
- **body**: Full textual description of the apartment and its features.

- **amenities**: List of available amenities, including items such as air conditioning, basketball courts, cable access, gym facilities, internet connectivity, pool access, and refrigerator availability.
- **bathrooms**: Number of bathrooms in the apartment.
- **bedrooms**: Number of bedrooms in the apartment.
- **currency**: The currency in which the price is originally listed.
- **fee**: Any additional fee associated with the apartment.
- **has_photo**: Boolean indicator for whether the listing includes a photo.
- **pets_allowed**: Specifies what types of pets (dogs, cats, etc.) are permitted.
- **price_type**: Standardized price in USD for comparative analysis.
- **square_feet**: The size of the apartment in square feet.
- **address**: Detailed address of the apartment.
- **cityname**: City in which the apartment is located.
- **state**: State in which the apartment is located.
- **latitude**: Geographical latitude of the apartment location.
- **longitude**: Geographical longitude of the apartment location.
- **source**: Source platform or website from which the listing was obtained.
- **time**: The timestamp when the listing was created.
- **price**: The numeric rental price of the apartment. This is the target variable.

The dataset is located at this URL.

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the apartment information for the development set. This portion has the "price" feature to be predicted for an apartment. This feature is the target value to be used to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the apartment corresponding to the evaluation set. This portion does not contain the "price" target variable.

2.2 Task

You are required to build a regression pipeline to predict the target variable, i.e., "price" columns of development.csv file.

2.3 Evaluation metric

Your submissions will be evaluated in terms of the Mean Absolute Error (MAE). Here you can find the function used to evaluate your submissions.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
0,650
1,820
2,1115
3,760
4,530
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `Id` in the evaluation CSV file.
- the `Predicted` float outcome. It must be in numerical form, as the development set provides.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_summer_2025**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in this form by the due date of this project. Failure to do so will result in a void project.

Warning: This is an additional requirement that was not required in past years. Make sure you remember to fill in the form by the due date, or your project will not be considered valid!