Summary of model and evaluation

The solution, together with some comments, can be found in *homework_notebook_caivano.ipynb* in a notebook format (or alternatively in *homework_plain_text_caivano.txt* in plain text format).
Data were shuffled on my local system through an external library written in C++ (https://github.com/alexandres/terashuf): all the files needed are attached in the folder "terashuf" (the only dependency should be stdlib).
The shuffling was performed in order to avoid the introduction of bias in the training of the model: data is more or less ordered by date in the original csv and this could introduce some bias.
The commands to perform the shuffling are (from the folder terashuf):
$ make
$ ./terashuf < pp-complete.csv > pp_complete_shuffled.csv
The model that I decided to use is a Neural Network (compiled in Keras) trained on the duration type, property type and town location (London or not) of the property.
Both duration type and property type are preprocessed using one-hot-encoding.
To deal with the dimension of the dataset I've imported the dataset through pandas in chunks: those chunks are iterated through a loop, in every step a chunk is splitted into train and validation (for samples corresponding to properties purchased before 01-01-2019) and test (samples corresponding to properties purchased after 01-01-2019) and the Neural Network is iteratively trained using the current chunk.
The Neural Network used is composed of 3 hidden layers on 10 neurons each, all with RELU activation functions. Moreover an Adam optimizer is adopted using mean squared error as loss function. Finally this model is used to predict on the test set and some evaluation metrics, such as mean absolute error and mean squared error, are computed and printed in order to evaluate how the model performs on the test set.
As we could expect, the predictions on the test set are not very accurate (probably because the model is very simple) with mean absolute error MAE=241219.74, mean squared error MSE=3411288110175.523 and root mean squared error RMSE=1846967.27.