ML4Science: The Green Building Project

Sébastien Houde (sebastien.houde@unil.ch)
Harald Mayr (harald.mayr@econ.uzh.ch)

Last update: November 14, 2022

1 Background

1.1 The Green Building Project

The Green Building Project is a collaboration between Sébastien Houde (HEC Lausanne) and Harald Mayr (University of Zurich). We work with leading companies in the Swiss real estate sector: a real estate owner, a real estate management company, and a real estate platform. Our research studies tenants' preferences for energy-efficiency and landlords' incentives to invest in energy-efficiency. More broadly, we are interested in the effects of energy prices, regulation, and information on the real estate market. The research is supported by an SNFS Ambizione grant and builds on insights from an earlier SNSF Spark grant.

1.2 Demand prediction for apartment listings

In this ML4Science project you will predict tenants' demand for apartment listings on online platforms. Your work will contribute to The Green Building Project in two important ways.

First, we plan to use your machine learning algorithm to increase the statistical power of planned field experiments. In our field experiments, we will vary the content and presentation of apartment listings. We want to estimate how these variations influence demand for apartments. As demand for apartments depends not only on the variations in our field experiment, but also on apartment characteristics (e.g., location, size, building standard), machine learning helps us to account for apartment characteristics and make listings comparable. Guo et al. (2021) describe the method.

Second, we plan to use your machine learning algorithm as an input for double machine learning (Belloni et al., 2014; Chernozhukov et al., 2018). The idea behind double machine learning is that machine learning can be used to control for variables that correlate with both the outcome variable (e.g., demand) and the variable of interest (e.g., price). A non-technical explanation of the method can be found here.

Your work can make an important contribution, because both approaches require a good prediction algorithm. This ML4Science project is very important to us. We will be happy to provide you with all the support you need to make it a great success.

2 Data

The data for this project contain 20,000 apartment listings that have been posted on a popular Swiss real estate platform. The file `listings.xlsx` contains the following columns:

```
['Are Pets Allowed', 'Category Idx', 'Day of Advertisement Created',
   'Day of Date Available From', 'Demand', 'Floor', 'Geo Canton',
   'Geo City', 'Geo Zip', 'Has Balcony', 'Has Cabletv', 'Has Elevator',
   'Has Fireplace', 'Has Garage', 'Has Parking', 'Is New Construction',
   'Is New Construction Potential', 'Is Tenant2Tenant',
   'Is Wheelchairaccessible', 'Listing Description', 'Listing Title',
   'Livingspace', 'Number Of Rooms', 'Number Of Rooms Cleaned',
   'Number of Documents', 'Number of Images', 'Price Extra Normalized',
   'Price Gross Normalized', 'Price M2 Normalized', 'Price Net Normalized',
   'Size M2 Normalized', 'Subcategory En Idx', 'Year Built',
   'Year Lastrenovated']
```

The most important column is `'Demand'`, which measures the number of users who filled out the online form to indicate interest in an apartment viewing. This is our outcome variable, the value we want to predict. The data set contains various features that may be useful to predict `'Demand'`:

- binary columns (e.g., `'Are Pets Allowed'`),
- categorical information (e.g., `'Geo Canton'`),
- cardinal information (e.g., `'Number of Images'`),
- dates (e.g., `'Day of Advertisement Created'`),
- and text (`'Listing Description'` and `'Listing Title'`).

Please note that `'Number of Images'` and `'Number of Documents'` provide only the number of images and documents that were uploaded. The image files are provided separately.

The data offers a few aspects that allow you to use sophisticated methodology. We are curious to see how you use natural language processing for the texts in `'Listing Description'` and `'Listing Title'`. Data on images and floor plans allow you to employ appropriate deep learning techniques (see e.g., Naumzik and Feuerriegel, 2020; Solovev and Pröllochs, 2021).

3 Prediction challenges

The aim of this project is to predict `'Demand'`. Your task is to train a machine learning algorithm to minimize out-of-sample loss, as measured by the mean Poisson deviance as implemented in

`sklearn.metrics.mean_poisson_deviance`. We keep a random hold-out sample to assess your predictions throughout the project.

Your algorithms will be evaluated based on three specialized challenges and one main challenge:

- 🏆 Best Overall: Prediction using all available data
- 🎗 Best Tabular: Prediction using only the tabular data (without texts)
- 🎗 Best Text: Prediction using only the texts in `'Listing Description'` and `'Listing Title'`
- 🎗 Best Image: Prediction using only image data

You are free to choose how you approach the prediction challenges. In particular, you may use any split of the data in training and evaluation sample, transformation of the data (i.e., feature engineering), machine learning algorithm (or combine different algorithms), and approach to cross-validation and tuning. This is a creative process and we encourage you to experiment and try different models and specifications. It is a journey, it should be fun, and you will gradually improve the model over time.

4 Timeline

The following timeline is tentative:

Date	**Task or event**
November 14	Kick-off meeting
Sprint 1	*Submit confidentiality agreements; study this document and course instructions*
November 21	Data access
Sprint 2	*Exploratory data analysis; feature processing; submit first predictions*
November 29	Progress meeting: Evaluation of predictions, discussion and questions
Sprint 3	*Refine algorithms; submit second set of predictions*
December 6	Progress meeting: Evaluation of predictions, discussion and questions
Sprint 4	*Finalize presentation; submit final predictions*
December 16	Final presentations; award ceremony
December 22	Submit report and code

All meetings will be virtual meetings (tentatively at 2pm). We will use the meetings for short progress updates, to address any questions you may have, and to share new data as it becomes available. If you have any questions, please do not hesitate to reach out in between meetings. Harald Mayr will be your main point of contact, but you may also contact Sébastien Houde.

References

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. (2021). Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, 34:8637–8648.

Naumzik, C. and Feuerriegel, S. (2020). One picture is worth a thousand words? the pricing power of images in e-commerce. In *Proceedings of The Web Conference 2020*, pages 3119–3125.

Solovev, K. and Pröllochs, N. (2021). Integrating floor plans into hedonic models for rent price appraisal. In *Proceedings of the Web Conference 2021*, pages 2838–2847.