Proceedings of Machine Learning Research 179:1–3, 2022 Conformal and Probabilistic Prediction with Applications

House Price Prediction with Confidence: Empirical Results from the Norwegian Market

Anders Hjort ANDERDH@MATH.UIO.NO
Department of Mathematics, University of Oslo and Eiendomsverdi AS

Editor: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

Abstract

Automated Valuation Models are statistical models used by banks and other financial institutions to estimate the price of a dwelling, typically motivated by financial risk management purposes. The preferred choice of model for this task is often tree based machine learning models such as gradient boosted trees or random forest, where uncertainty quantification is a major challenge. In this empirical contribution, we compare split conformal inference, conformalized quantile regression and Mondrian conformalized quantile regression on data from the Norwegian housing market, and use random forest as a point prediction. The data consists of $N = 29\,993$ transactions from Oslo (Norway) from the time period 2018-2019. The results indicate that the methods using conformalized quantile regression create narrower confidence regions than split conformal inference.

Keywords: Conformal inference, conformalized quantile regression, automated valuation models.

1. Introduction

Automated Valuation Models. Automated Valuation Models (AVMs) are models used by banks or other financial institutions to get an assessment of the estimated value of a dwelling. The most common model for this is a hedonic model that estimates the price based on the dwellings' attributes. While this historically has been a linear regression model (Bailey et al. (1963)), many recent studies indicate that machine learning models such as gradient boosted trees or random forest often have better prediction accuracy (Sing et al. (2021), Kim et al. (2021), Hjort et al. (2022)).

Conformal prediction. Conformal prediction (CP) is a model-agnostic framework for uncertainty quantification. The distribution of the absolute residuals $|y_i - \hat{y}_i|$ is used to form confidence regions for new and unobserved instances. One extension of the original CP framework is to construct a *Mondrian* CP (Shafer and Vovk (2007)), where the feature space is split into a set of non-overlapping categories and confidence regions are created separately in each category. Another recently proposed extension by Romano et al. is *conformalized quantile regression* (CQR), that combines the idea of conformal prediction intervals with the quantile regression framework.

Bellotti (2016) and Bellotti and Lim (2021) are both using conformal inference to create confidence bands for AVMs with applications to the UK housing market. The literature is otherwise quite sparse in terms of studies of uncertainty quantification in AVMs.

Table 1: The results of various CP methods applied to the data set of transactions from Oslo. The interval sizes are given in million Norwegian kroners (NOK). 1 NOK $\approx$ 0.1 USD per July 2022.

Method	Coverage (%)	Mean interval size	Median interval size
Split CP	89.54	1.85	1.61
CQR	90.25	1.79	1.23
Mondrian CQR	88.45	1.49	1.13

2. Preliminary results

This paper studies a novel data set of $N = 29\,993$ transactions from Oslo, the capital of Norway, from the two year period $2018 - 2019$. The mean sale price in the data is 4.7 million kroners. We use random forest as a point prediction with 500 trees, each with a max depth of 10. The sale price is predicted based on a total of $p = 13$ covariates such as size of the dwelling (measured in m^2), the number of bedrooms, the coordinates of the dwelling and neighborhood characteristics. We then experiment with different ways of creating confidence bands around the predictions. Normalized split CP; CQR and a Mondrian CQR set up where we utilize the 15 different city districts in the data set and create confidence regions with CQR in each city district.

The Root Mean Squared Error (RMSE) of the random forest point prediction is 11.9%. A comparison of the CP methods used to create confidence regions valid at $\alpha = 0.1$ on the test set can be seen in Table 1. The split CP method use $\sigma(\cdot) = \exp\{\gamma \cdot \hat{\mu}(x_i)\}$, where $\hat{\mu}(x_i)$ is a GAM model fitted on the residuals from the training set and γ is a hyper parameter. This follows the notation from Bellotti and Lim (2021). The CQR methods use the `quantregForest` package to create the quantiles.

The results indicate that CQR create narrower confidence regions (measured in Norwegian kroners) than normalized split CP. The Mondrian CQR achieves the narrowest intervals, but achieve an empirical coverage of 88.45%, indicating too narrow intervals.

3. Future work

A major challenge in house price prediction tasks it the temporal dimension; house prices change over time as a result of market movements. This also affects the task of creating confidence regions, as we also would expect these to change with time. A natural direction for future research is to utilize the growing literature on conformal inference for time series, for instance by building on the works of Xu and Xie (2021).Another enticing option is to create confidence bands that account for covariate shifts, as outlined in Tibshirani et al. (2019). This is particularly relevant to our application, as different segments of the housing market tend to have very different characteristics. For instance, the dwellings in City A might be quite different from the dwellings in City B. It will be very useful to investigate how we can achieve valid and efficient confidence regions by accounting for this shift in distribution for some or all of the characteristics of a dwelling.

Acknowledgments

Thanks to Eiendomsverdi AS for providing the data. I am also grateful to Johan Pensar, Ida Scheel and Dag Einar Sommervoll for fruitful discussions about the topic.

References

Martin J. Bailey, Richard F. Muth, and Hugh O. Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304): 933–942, 1963.

Anthony Bellotti. Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 81:71–84, 2016.

Anthony Bellotti and Zhe Lim. Normalized nonconformity measures for automated valuation models. *Expert Systems with Applications*, 180:115–165, 2021.

Anders Hjort, Johan Pensar, Ida Scheel, and Dag Einar Sommervoll. House price prediction with gradient boosted trees under different loss functions. *Journal of Property Research*, 0(0):1–27, 2022. doi: 10.1080/09599916.2022.2070525.

Jungsun Kim, Jaewoong Won, Hyeongsoon Kim, and Joonghyeok Heo. Machine-learning-based prediction of land prices in seoul, south korea. *Sustainability*, 13(23):202–211, 09 2021. doi: 10.3905/jpm.2017.43.6.202.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 07 2007. doi: 10.1145/1390681.1390693.

Tien Foo Sing, Jesse Jingye Yang, and Shi Ming Yu. Boosted tree ensembles for artificial intelligence based automated valuation models (ai-avm). *The Journal of Real Estate Finance and Economics*, 43:202–211, 09 2021. doi: 10.3905/jpm.2017.43.6.202.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. In *Advances in Neural Information Processing Systems*, 2019.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR, 18–24 Jul 2021.