Deep learning for real estate price prediction

By Walthert, Lorenz and Sigrist, Fabio, Deep Learning for Real Estate Price Prediction (May 24, 2019). Available at SSRN: https://ssrn.com/abstract=3393434 or http://dx.doi.org/10.2139/ssrn.3393434

Paper Review: By Krishna Jha

Introduction

Pricing of real estate objects is a fundamental problem in finance, urban development, and politics. Specifically state like California and Texas for investment in the housing market. Usually, the cities with high property value are also the one with the best location in terms of school rating, connectivity, low crime rate, public transport, and walkability etc. Since such cities are in high demand and always sought after, they form golden clusters of cities for the purpose of investment and good ROI as well as for residential purposes.

Objective & Data

The competition goal is to predict sale prices for homes Switzerland based on model like gradient boosted regression Trees. This research use a data set of transactions of Swiss residential apartments. The data contain 88’247 samples ranging from the first quarter of 2011 to the end of 2017. For each transaction, the price of the property is recorded along with several attributes of the property. In this study mostly data analyses as a data dictionary. Capabilities of deep learning for real estate appraisal and compared it with a linear model as well as gradient boosted regression trees. This study conducted an extensive network architecture search and evaluated four main classes of deep learning models which are characterized by their layer composition.

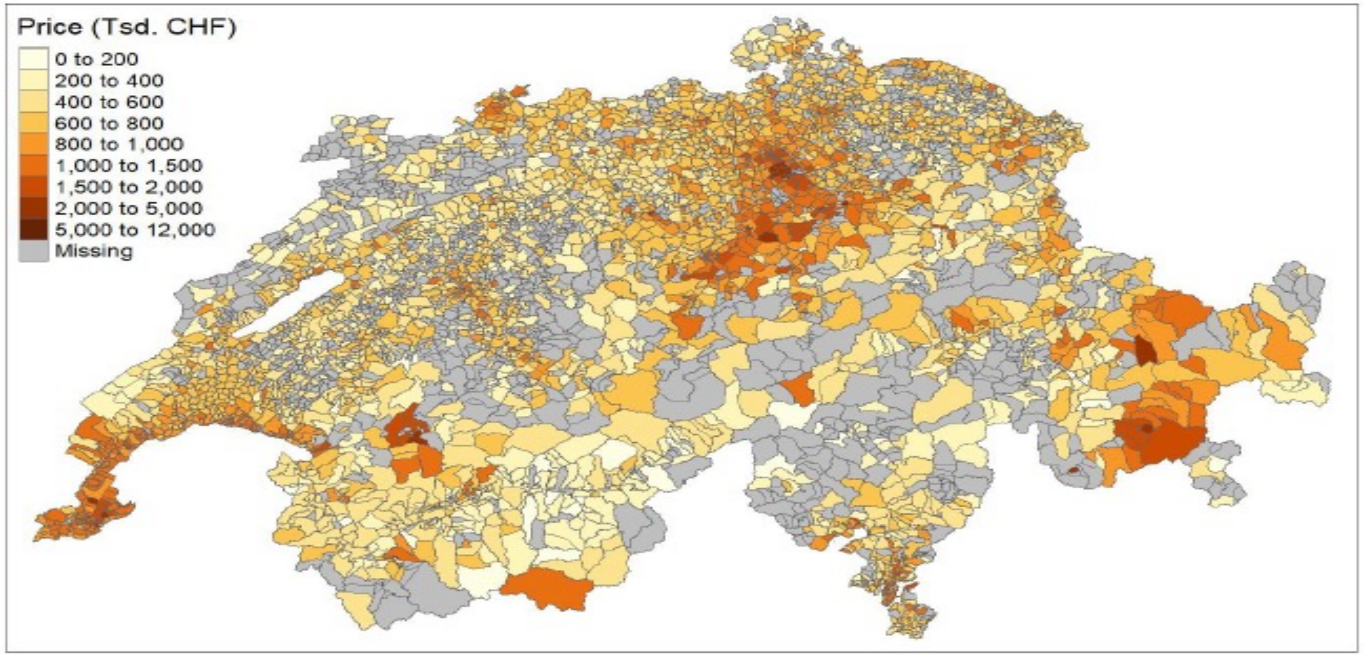
Data description and visualization

Table 1: Description of variables available in the raw data set.

attribute name	attribute description	attribute type	n levels
price	transaction price	numeric	
district	district / municipality	factor	3471
constyear	year of construction	integer	
nbroom	number of rooms	numeric	
cond	condition	factor	9
stand	standard	factor	9
micro	micro location rating	factor	9
renov	difference between theoretical and actual renovation standard	factor	9
holiday	vacation apartment indicator	factor	2
holiday2	interaction vacation apartment indicator and vacation apartment law	factor	6
area	main usage area according to SIA 416	integer	
park	additional parking lots	integer	
quarter	quarter of transaction	integer	
type	type of condominium / level	factor	13

Table 2: Summary statistics for relevant variables.

variable	mean	sd	min	pctl(25)	pctl(75)	max
price	705’167.500	502’472.700	56’000	410’000	845’000	11’850’000
constyear	1’995.158	20.516	1’920	1’984	2’010	2’017
nbroom	4.077	1.184	1.000	3.500	4.500	20.000
cond	7.458	2.077	1	7	9	9
stand	5.695	1.404	1	5	7	9
micro	6.220	1.230	1	5	7	9
renov	0.256	0.638	0.000	0.000	0.000	4.000
area	111.638	40.186	20	85	133	350



Deep learning Methodology

- Linear regression
- Meta model
- Gradient boosting
- Deep learning

  -> Neurons

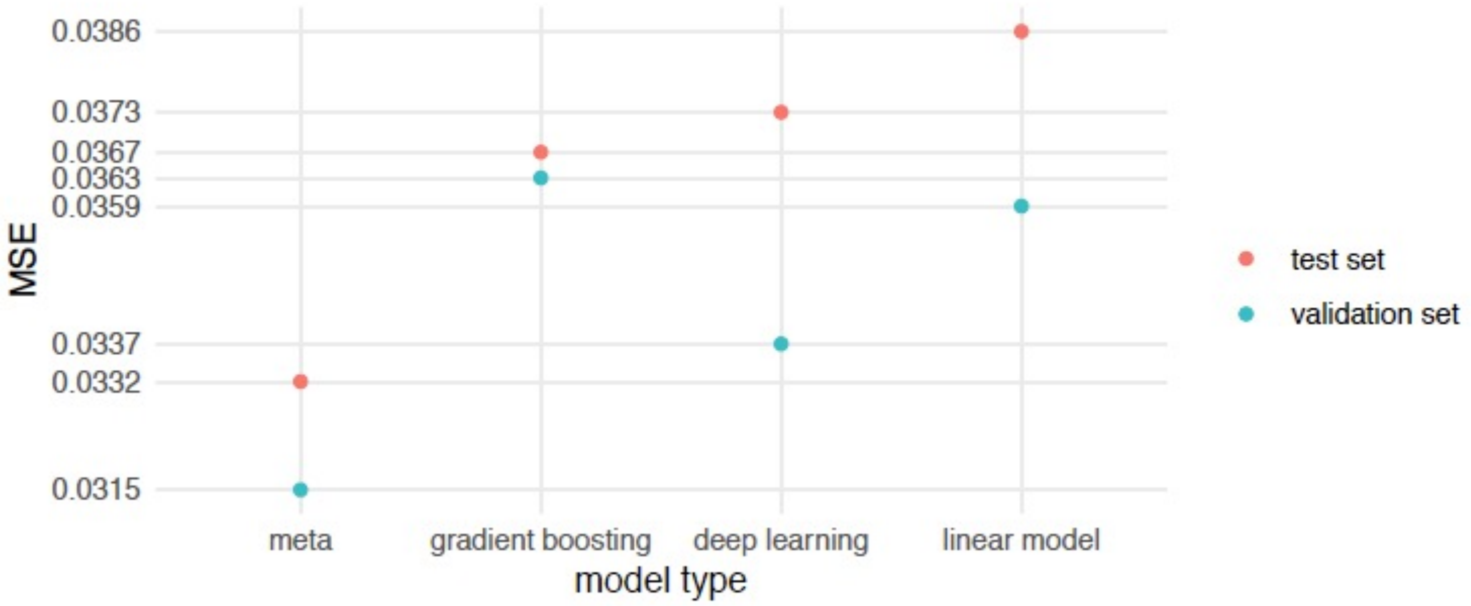
  -> Pyramid structure

  -> Flat structure

To evaluate the predictive performance of the models, we use the mean square prediction error (MSE) of the logarithmic transaction price.

Results

The MSEs on the out-of-sample test data for the different models are shown in below Diagram. In addition, research show the MSEs on the validation data. The linear model achieves a test MSE of 0.0383, boosted trees an MSE of 0.0367, deep learning an MSE of 0.0373, and the meta model an MSE of 0.0332. I.e., the meta model achieves the highest predictive accuracy. The second-best model is gradient boosted trees. Study observe almost no difference between the validation and the test MSE. The figure also shows that there is possibly some over-fitting occurring for the deep learning model as the difference between the validation and test MSE is relatively large. Finally, the linear model achieves the lowest predictive accuracy.



The Estimated fixed-effects coefficients

Test results for comparing the different models. The tests are done using a regression model with a random effect at the observation level.

- t- and p-values are reported in below table. These results indicate that the linear model is significantly worse than the deep learning model and that the difference between gradient boosting and deep learning is not significant at the 5% level. In addition, the meta model is clearly superior to the deep learning model.

	Estimate	t-value	Pr(>|t|)
(intercept)	3.7E−02	5.7E+01	0.0E+00***
gradient boosting	4.1E−01	3.1E−04	6.8E−01
linear model	1.6E−03	5.0E+00	5.4E−07***
meta model	−3.8E−03	−1.3E+01	0.0E+00***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Conclusions

This research investigated the capabilities of deep learning for real estate appraisal and compared it with a linear model as well as gradient boosted regression trees. This Study conducted an extensive network architecture search and evaluated four main classes of deep learning models which are characterized by their layer composition as well as by how **spatial effects** are incorporated in the model.

Each model class, study considered several tuning parameters such as the number of hidden layers, units per hidden layer, regularization, and the learning rate. The results show that a neural network model performs significantly better than a linear regression model, although the difference is economically small. The meta model, which averages the predictions of linear regression, gradient boosting, and deep learning achieves a considerably and significantly better performance than all other models.

References

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3393434
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques