Predicting and Mapping Danish House Prices

Daniel Thorsten Bruns mws487, Moritz Bilstein wmj863

Introduction to Social Data Science

University of Copenhagen

Assignment of sections:

- Introduction: mws487
- Literature Review: mws487
- Data: wmj863
- Visualisation: mws487
- Prediction Model: wmj863
- Discussion: Both
- Conclusion: wmj863

Copenhagen, August 23, 2022

Contents

List of Figures

List of Acronyms

API Application Programming Interface

CRS Coordinate Reference System

OSM OpenStreetMaps

RMSE Root Means Square Error

1 Introduction

Predicting house prices is not only important to individual buyers and sellers, but also for the establishment of real estate policies. Furthermore, house prices are an important factor in determining the general state of the economy. There are many use cases for unbiased models to predict house prices. The literature has established house characteristics like lot size or age and location features as the set of important price determinants. This study obtains these information for the current danish housing market and builds a prediction model using a lasso regression approach. The machine learning approach allows us to ask which set of features is most important to predict house prices in Denmark. How accurately can the asking price for a house be determined only by considering it's characteristics? For the analysis we focus on currently listed prices of available properties on boligsiden.dk. We evaluate the predictive accuracy using the Root Means Square Error (RMSE) and find that the location of houses in our dataset are the most important set of attributes determining price. We further visualise spatial price differences in a heatmap obtained from OpenStreetMaps (OSM).

The paper continues in the following structure. We first present previous studies that shaped a trend towards machine learning methods. Second, we describe how we obtained the dataset and present geographical visualisations. Third, we introduce our theoretical framework and present our results. We end by discussing our results and approach and concluding.

2 Literature Review

While hedonic-based models have been dominating the literature on house price predictions, recent studies utilize machine learning algorithms to overcome limiting assumptions. Potentially limiting features of the traditional approach are assumptions of independent variables and functional forms (Selim 2009). Various studies compare hedonic with machine learning methods and find generally improved prediction accuracy for the latter (Kauko, Hooimeijer, and Hakfoort 2002; Liu, Zhang, and Wu 2006; Se-

lim 2009). Park and Bae (2015) compare four different machine learning algorithms to model housing prices in Fairfax County, Virginia to enhance predictive accuracy. Manasa, Gupta, and Narahari (2020) use lasso and ridge regression models to predict urban housing prices. The recent trend towards machine learning based methods and away from conventional statistical approaches motivates our study, which is among very few applications to the danish residential market.

3 Data

3.1 Data Acquisition

The house price data that we work with is obtained on boligsiden.dk. We download json files from the boligsiden Application Programming Interface (API) and convert them into a dataframe. We request 20 batches of 500 real estate sale advertisements each to obtain data on the 10,000 most recently added advertisements. Time and date of the requests is stored in a log-file. There are feasible alternatives to boligsiden.dk like boliga.dk and nybolig.dk.

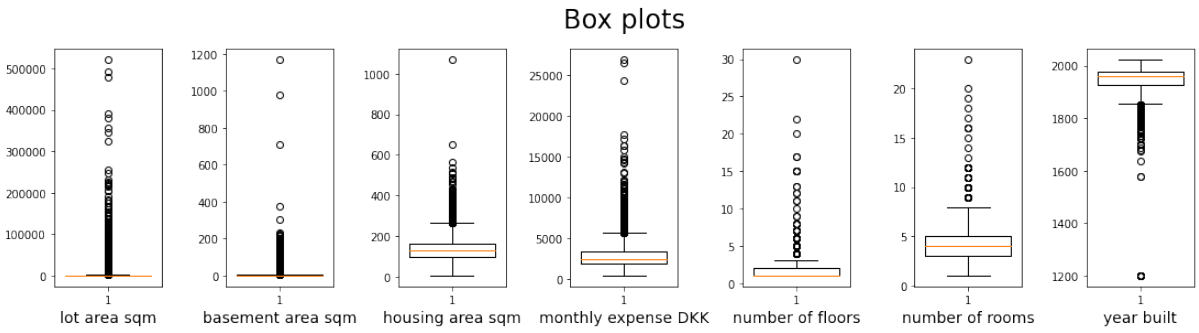
3.2 Data Structuring

The restructuring of the json files is straightforward, as each advertisement can be accessed individually in each json file. We convert the data into a pandas dataframe so that the characteristics of the advertisements can be found in the rows, and the respective variables in the columns. We make sure that the information in the cells have the right type, so that continuous numbers are stored as floats, integers as integers and so forth.

We focus on variables describing the real estate units themselves, so we remove variables concerning the advertisements, like date and type of advertisement and viewer reactions. We end up with variables on basement, housing and lot area, capturing the monthly expense, number of floors, number of rooms and the year the unit was built. Figure 1 illustrates the distribution of the non-indicator data through box plots. They

show the wide variety of of the houses. So does the majority of houses not have an attached lot at all, but many have lots with sizes in their thousands of m^2. The median house does not have a lot, has no basement, a housing area of 131 m^2, monthly expenses of 2561 DKK, one floor, four rooms and was built in 1973.

Figure 1: Box plots of the non-indicator covariates



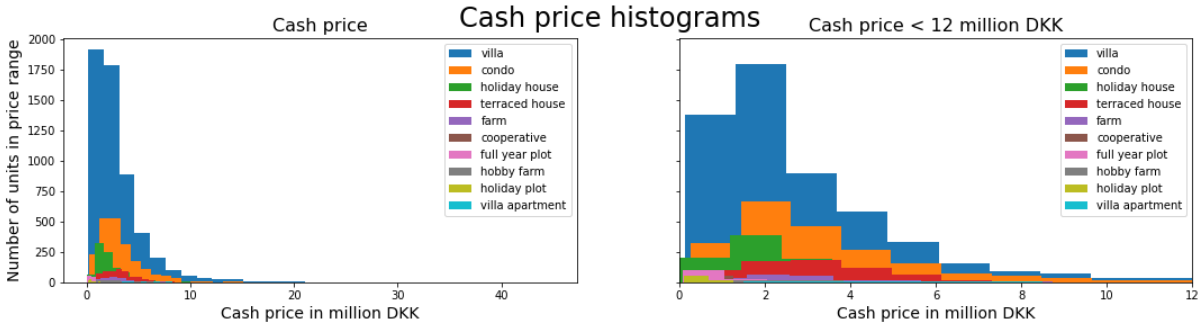
Notes: The figure shows the seven variables referred to as non-indicators in results tables 1 and 2.

Additionally, we create indicator variables for the type of real estate (villa, condo, farm, etc.) the energy label (g, f, ..., a, a2010, a2015, a2020) and the municipalities.

The energy label variable has many missing observations - 1299, which is more than 10% of the data. We have two options. We can either drop the advertisements with missing values and continue with a reduced data set or drop the variable. The machine learning models will show that the energy label is an important determinant of the house price, so we decide to keep the variable and drop the advertisements which do not report an energy label. Table 5 shows the occurrences of energy labels.

We have data on nine different house types which can be seen in table 4 in the appendix. Figure 2 shows the cash price distributions of the houses depending on the type. Houses of the type *villa* are most frequent in every price category.

Figure 2: Histograms of the cash price distribution depending on address type



Notes: Left: All observations, Right: Zoomed in on observations below a cash price of 12 million DKK

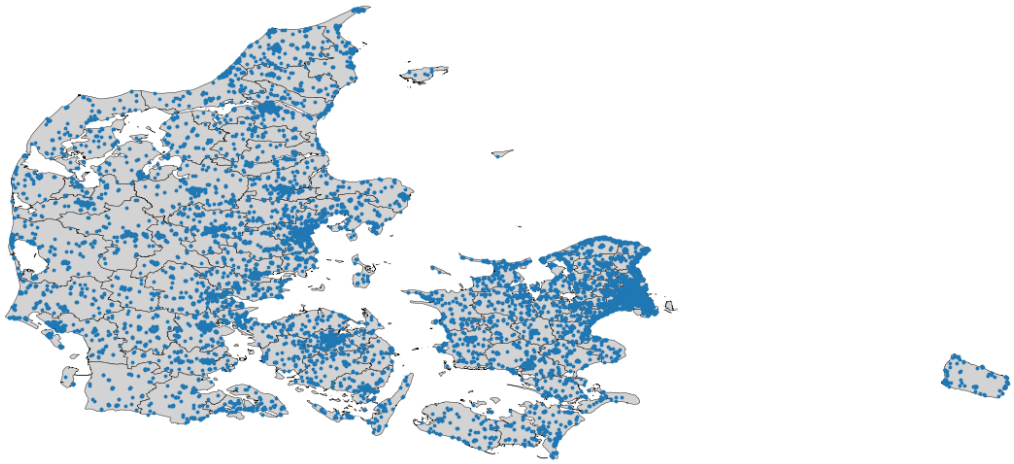
All 98 municipalities are represented in our data set. The municipality with the least observations is Læsø (5) and København (758) has the most observations. We list all absolute frequencies across municipalities in table 6 in the appendix.

Ultimately we end up with a clean data table of 8672 advertisements with twelve characteristics. For the analysis we transform the energy labels, address types and municipality labels into indicator variables, so that we end up with a table of 8672 advertisements and 129 variables.

4 Visualisation

We use spatial visualisation for further investigation of our data set. OSM is an open source map that allows the user to produce personalized maps for any purpose. Furthermore, it provides an API, which we use to request the spatial information of administrative boundaries in Denmark on municipality level. After converting both the obtained json file and out main data set into a geopandas dataframe, we align the Coordinate Reference System (CRS) in both dataframes. We collapse our main data set that contains housing characteristics to the municipality level and merge it with the geographical information. Plotting the geographical locations of our observations and housing characteristics gives us a better feel for the data set at hand and its real world application.

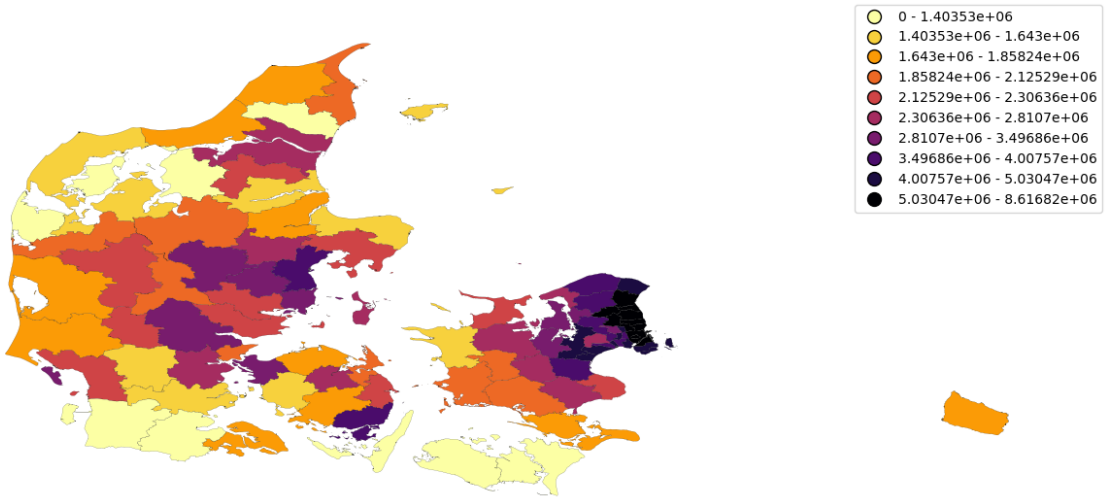
Figure 3: Spatial Distribution of Sample across Denmark



Notes: The figure shows all listed houses contained in the final data set.

Figure 3 presents all houses contained in our sample. It confirms that our data represents all 98 danish municipalities and gives a again the impression that house listings are more frequent in the urban areas of Denmark.

Figure 4: Average House Prices across Municipalities, Denmark



Notes: The figure shows average house prices on municipality level. Darker colors indicate increasing prices in DKK.

The heatmap in figure 4 helps us to establish the importance of properties' locations in determining housing prices. It presents differences in average housing prices across

danish municipalities ranging from $1,018,406$ DKK in Morsø to $8,616,821$ DKK in Gentofte. The stark differences give a first indication to the importance of location as a feature in the prediction of housing prices. Furthermore, the map confirms our intuition that the urban areas around the Denmark's largest cities are most among the most expensive areas.

5 Prediction Model

Machine learning algorithms are the state of the art methods for predictive tasks. They especially thrive for issues that involve a lot of data. Our raw data set contains 8672 observations with 128 covariates, regular statistical methods can be overcharged by the amount of variables, especially when the covariate set grows in interaction and polynomial terms. Hence machine learning is the ideal toolbox for our problem to predict house prices given their characteristics.

Our prediction problem is conceptionally continuous so a form of regression is called for. The lasso is a regularizing linear regression model that automatically performs variable selection.

$$\hat{\beta}_{\lambda}^{L} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

To shrink the model parameters penalty term of λ is added to the ordinary least squares minimisation formula. The parameters of β that are closest to zero are then forced to be zero.

We split the data into 80% training and 20% test data. To determine the optimal value for λ we use five fold cross validation on the training data. Other methods are the Bickel-Ritov-Tsybakov rule, the Belloni-Chen-Chernozhukov-Hansen rule which find a feasible penalty term analytically or search methods like grid and random search. It is important to note that we standardise our data before determining λ. The indicator variables are not standardised. And the training and test data are standardized independently from each other, so that no information of the test data corrupts the training

Results with polynomial order 2

Data set	Selected Var	Total Var	R^2	RMSE	λ
non indicators	7	7	0.6350	1803552	3963
poly interaction	21	35	0.6854	1744585	10527
address indicators	32	46	0.7232	1687382	6926
energy indicators	30	47	0.6967	1724654	11288
municipality indicators	122	133	0.7683	1523098	2268
all indicators	133	156	0.7875	1484578	4557

Table 1: Results of the lasso from row two with interaction terms and polynomials of order two. Indicator data sets add certain indicators to the data with interaction terms and polynomials

Results with polynomial order 4

Data set	Selected Var	Total Var	R^2	RMSE	λ
non indicators	7	7	0.6350	1803552	3963
poly interaction	36	329	0.7158	1782323	13917
address indicators	42	340	0.7451	1684571	12979
energy indicators	41	341	0.7293	1742927	13917
municipality indicators	112	427	0.7843	1497692	12979
all indicators	125	450	0.8048	1466311	12979

Table 2: Results of the lasso from row two with interaction terms and polynomials of up to order four. Indicator data sets add certain indicators to the data with interaction terms and polynomials

data and with that our estimates.

5.1 Results

We estimate the lasso in several different ways and use one metric each to assess in-sample (R^2) and out-of-sample performance RMSE. The RMSE is computed based on the model predictions for the test data. We start with using only the seven non-indicator variables. Then we add polynomial terms and interaction terms. Additionally to the polynomial and interaction terms we add one set of indicator variables each, to examine their importance for the prediction performance. This we do twice, once with polynomials of a maximum order of two and once with a maximum order of 4. The results can be found in tables 1 and 2. The tables let us reconstruct what variables are how important.

We see that municipality indicators increase the R^2 and decrease the RMSE the most, followed by the address type indicators and then the energy labels. In both cases all the indicator variable sets are more important than the polynomials and interaction terms of the non indicator variables. The polynomials of up to order four deliver a better in-sample performance compared to the polynomials of up to order two, as the R^2 is higher by around 0.03 points. The out-of-sample performance is worse though, with a higher RMSE. In general, the in-sample fit is always better for the models with higher polynomials, because the R^2 can only increase in more variables. The best performing model has polynomials of up to order four and contains all indicator variables. It performs best in both metrics and reaches an R^2 of 0.8048 and an RMSE of 1466311.

6 Discussion

Our lasso model performs best in both in the in-sample and out-of-sample metric, so we can be sure it did not over fit in the training data and that the cross validation to determine λ worked. The R^2 is also not inflated by the number of coefficients, as the R^2_{adj} is 0.802. The RMSE is unfortunately economically high and is assumed to not be compatible to professional valuation of houses. Building upon that, further research should include an analysis if there are certain houses the machine learning algorithm can predict well while the RMSE is influenced by certain types that the algorithm cannot valuate as well. A hypothesis that the RMSE would be far lower if we made a cut on the cash prices and disregarded houses from a certain cash price of e.g. 10 million DKK is not backed up by table 3. For this table we repeated the analysis for this sub sample.

There might be a potential bias from using the most recent advertisements posted to boligsiden.dk. Houses that do not sell for the high asking price might have been reposted. Park and Bae (2015) address this difference and try to predict the difference between listing and actual closing price of properties.

It is important to note that RMSE has alternatives like the Mean Average Error or the Mean Relative Error that could have been investigated additionally or alternatively. The

Mean Relative Error would mitigate larger errors for houses with higher prices.

There are alternative regularization methods for our linear regression model like ridge and elastic net that were not analysed. Additionally, there are other machine learning approaches like tree-based learning or support vector machines. A reasonable extension would be to extend the interaction terms to the indicator variables. However, the increased computation time would not have fit to the time scope of this paper.

7 Conclusion

In this study, we use a lasso regression approach to determine the most important set of features in predicting house prices. We find that location of listed houses is the set of features that increases predictive accuracy the most. We find that living in one of the urban areas of Denmark increases average house prices substantially. Our final model has a very high in-sample fit, but a very high out-of-sample fit. The advantages of lasso are found in its simplicity and narrow design of a singular estimation equation. Other machine learning algorithms should be considered for further investigations.

Upon further extensions, a predictive model could bypass the manual valuation of properties based on individual experience and knowledge of brokers and inform real estate policy that depends on predictions of future house prices. Future research could apply additional machine learning algorithms to enhance predictive accuracy of the model. Furthermore, future research could integrate additional inputs such as environmental amenities and macroeconomic variables that potentially affect house prices.

References

Kauko, T., P. Hooimeijer, and J. Hakfoort (2002). “Capturing housing market segmentation: An alternative approach based on neural network modelling”. In: *Housing Studies* 17.6, pp. 875–894.

Liu, J.-G., X.-L. Zhang, and W.-P. Wu (2006). “Application of fuzzy neural network for real estate prediction”. In: *International Symposium on Neural Networks*. Springer, pp. 1187–1191.

Manasa, J., R. Gupta, and N. Narahari (2020). “Machine learning based predicting house prices using regression techniques”. In: *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*. IEEE, pp. 624–630.

Park, B. and J. K. Bae (2015). “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data”. In: *Expert systems with applications* 42.6, pp. 2928–2934.

Selim, H. (2009). “Determinants of house prices in Turkey: Hedonic regression versus artificial neural network”. In: *Expert systems with Applications* 36.2, pp. 2843–2852.

A Appendix

Results with polynomial order 2 for cash price < 10 mio DKK

Data set	Selected Var	Total Var	R^2	RMSE	λ
X 7	6	7	0.5614	1312557	13040
poly interaction	17	35	0.6256	1257664	6053
address indicators	34	46	0.6830	1159684	2279
energy indicators	30	46	0.6515	1238588	4270
municipality indicators	121	133	0.7580	1110323	1398
all indicators	140	155	0.7940	1030174	1398

Table 3: Results of the lasso from row two with interaction terms and polynomials of order two. Indicator data sets add certain indicators to the data with interaction terms and polynomials. Sub set of advertisements with an asking price below 10 million DKK

Housing type absolute frequencies

address type	number of observations
condo	2025
cooperative	150
farm	217
full year plot	8
hobby farm	101
holiday house	26
terraced house	669
villa	5413
villa apartment	63
basement	2189
lot	6579

Table 4: Absolute frequencies of housing types

Energy label absolute frequencies

energy label	number of observations
a	1
a1	1
a2	5
a2010	271
a2015	301
a2020	137
b	626
c	2872
d	2694
e	991
f	523
g	250

Table 5: Absolute frequencies of energy labels

Municipality absolute frequencies

municipality	number of observations
Aabenraa	59
Aalborg	313
Aarhus	508
Albertslund	16
Allerød	25
Assens	67
Ballerup	45
Billund	29
Bornholm	98
Brøndby	17
Brønderslev	83
Dragør	31
Egedal	74
Esbjerg	134
Faaborg-Midtfyn	75
Fanø	6
Favrskov	66
Faxe	84
Fredensborg	65
Fredericia	55
Frederiksberg	126
Frederikshavn	111
Frederikssund	61
Furesø	60
Gentofte	132
Gladsaxe	67
Glostrup	41
Greve	70
Gribskov	68
Guldborgsund	106
Haderslev	97
Halsnæs	47
Hedensted	78
Helsingør	124
Herlev	21
Herning	87
Hillerød	77
Hjørring	132
Holbæk	146
Holstebro	79
Horsens	125
Hvidovre	59
Høje-Taastrup	86
Hørsholm	56
Ikast-Brande	54
Ishøj	31
Jammerbugt	78
Kalundborg	74
Kerteminde	27
Kolding	144
København	758
Køge	119
Langeland	29
Lejre	41
Lemvig	32
Lolland	66
Lyngby-Taarbæk	88
Læsø	5
Mariagerfjord	79
Middelfart	47
Morsø	28
Norddjurs	62
Nordfyns	55
Nyborg	37
Næstved	179
Odder	47
Odense	257
Odsherred	69
Randers	195
Rebild	47
Ringkøbing-Skjern	71
Ringsted	51
Roskilde	138
Rudersdal	125
Rødovre	61
Samsø	9
Silkeborg	168
Skanderborg	77
Skive	52
Slagelse	138
Solrød	46
Sorø	41
Stevns	44
Struer	38
Svendborg	84
Syddjurs	79
Sønderborg	75
Thisted	64
Tårnby	69
Tønder	52
Vallensbæk	23
Varde	83
Vejen	77
Vejle	154
Vesthimmerlands	66
Viborg	137
Vordingborg	77
Ærø	19

Table 6: Absolute frequencies of municipalities