Predicting apartment prices in the urban area of Copenhagen
A machine learning approach

Sarah Koelemij Andersen, Mathias Nichlas Gottschalck,
Mikkel Nørholm Jansen, Thor Noe

31st August 2018

Abstract

In this paper we try to predict listing prices on owner-occupied apartments in the urban area of Copenhagen. We use data scraped from web pages and include different attributes such as distance to top tier schools, metro stations and jail buildings. We use different machine learning approaches and benchmark the results against a classic linear OLS regression. Our results show that a Lasso cross validated model with 3 polynomial features gives the best predictions. However, we get a prediction error of 12 percent, corresponding to an error of DKK 0.5 million for an average apartment of 93 square meters. Though our model does not predict better, machine learning could potentially be part of future house valuation for both professional real estate agents and common homeowners.

Contributions by exam number

44: 3.1, 4.1, 5.1

59: 3.2, 4.2, 5.2

124: 3.3, 4.3, 6.1

209: 3.4, 4.4, 6.2

Keystrokes (normal pages): 40,000 (17)

Abstract, Introduction, Related literature, Discussion, Conclusion and all research were prepared jointly. Our results are fully reproducible by running the Jupyter Notebook file.[1]

[1] https://github.com/thornoe/sds_2018/blob/master/CPH/Notebook_final.ipynb

Contents

1 Introduction

Selling a home is a big decision. Not only does one have to say goodbye to a lot of memories, but there are also many practical considerations that have to be taken into account. In this paper, we will investigate if machine learning methods can be used to estimate owner-occupied apartments listing prices in the urban area of Copenhagen. To do this, we extract data from Bolighed.dk, as well as other sources to analyse whether these new methods can estimate the same price as real estate agents. If it is possible, it could potentially be used to increase competition on this market and press down fees which typically range between 50.000-100.000 DKK, not including stamp fees, Land Registry certificate, etc [Bolius, 2018]. Also, these methods can be utilised by common people to list sales without an agent by providing address as well as house characteristics. The gain is obviously the saved fee the real estate agent would otherwise have taken (stamp fee's, Land Registry certificates, etc., are still required).

Setting the right price is important. The seller want the price to be as high as possible to maximise their gain. But usually they also want the house to be sold quickly. Firstly, because the residents wants to move on quickly with their lives. Secondly, there is a negative correlation between time-on-market and real estate prices [Stacy Sirmans, 2010].

There are different approaches to valuate a home. One is to compare the house with equivalents in the market, but since the market for housing is less liquid and has a smaller turnover than for e.g., financial products, this is not straightforward. One can instead use the *net income approach* where the potential rental income during a year is summed and divided by the estimated growth rate in this business. Lastly, The house can be valued using a *discounted cash flow model* where all future cash flows are discounted back to present value.

These approaches are all demanding and is more commonly used when buying commercial properties or property investment. It is not realistic to assume that real estate agents use these methods. Instead, they rely on the price level of the neighbourhood, the amount of rooms, the square meters, distance to top tier schools, possibility of public transportation, and grocery opportunities. These are some of the factors regional director Matthew Cooke [2016] proclaims they look at in the valuation process.

In this paper, we examine whether we are able to predict the same price as a real estate agents only using a machine learning approach on the publicly available data. That is, *our goal is to predict housing prices that are close to the real estate agent's valuation.*We are aware that we predict only the supply prices and not the true market prices as we reproduce any systematic pricing mistakes of the real estate agents. If we are successful, and learn that real estate agents are redundant, house seller's suddenly have another DKK 50,000-100,000 to spend on a new car, vacation, etc. It is worth mentioning that by selling without an agent you will have to bear all the risk that follows from selling a house by yourself which might be a burden for some risk averse individuals. Also, seller would have to be in charge of all practicalities, such as showcasing the apartment, take pictures and find potential buyers, which takes time and increase the opportunity costs of using an agent.

Using cluster analysis and visualization in GeoPandas we see that location has potential to explain much of the variation in price, thus, we collect several location variables besides the data on the apartments.

Our best prediction model use Lasso cross validation, where we use K-fold cross validation to optimise over λ. Our mean prediction error is 12.3 percent which for an average apartment in our sample accumulates to DKK 493,176 off from the target price with an actual listing price on DKK 4.1 million. Thus, we are not able to estimate listing prices precise enough to discard real estate agents. However, based on the methods we test on our relatively small sample, we cannot reject that more advanced machine learning methods, or even just a larger data sets with more variables, could make the real estate agent redundant.

The rest of the paper is structured as follows: Section 2 provides a short overview of related literature on predicting house prices. Section 3 explains how and where we scrape data. Section 4 presents a descriptive overview and cluster analysis. Section 5 goes through the implementation of machine learning and how it performs relative to Ordinary Least Square. In Section 6 we presents our results. In Section 7 we discuss our result and, finally, in Section 8 we conclude our findings.

2 Related literature

There is a prolific literature on predicting house prices in the field of economics. A popular approach is to use an *error correction model* (see, among others, Gattini and Hiebert [2010]; Ester Hansen and Staghøj [2013]), which is useful for estimating both short- and long-term effects. Another approach is to use the classic linear regression and *ordinary least square* (OLS) for estimation. However, both methods are subject to either under- or over-fitting, i.e. results in a model that either is too simple and does not capture the true relationship in data, or is too complex and captures all variation in the sample, but then performs poorly in out-of-sample prediction.

In this paper, we use machine learning to accommodate some of these challenges, and to improve upon the prediction power of our model relative to OLS. This is in line with Masías et al. [2016]. They use market data on house prices in Chile and find that machine learning models, such as *Neural Network* (NN), *Random Forest* (RF) and *Support Vector Machine's* (SVM) all perform better than OLS. Our paper differs from their work, in that we do not want to predict the true market price. Instead, we will predict the valuation set by real estate agent's and see whether these machine learning methods can predict the same or similar prices. If that is the case we will be able to cast light on new methods that could reduce the cost or potentially replacing their fee by using new methods for valuation.

3 Scraping, variable selection and cleaning

In this section we go through how we collect and clean our data. Most of the data is scraped from Bolighed [2018], but we also use coordinates for schools, metro stations and jails. We also go through the ethics of scraping the data. Furthermore, we construct a variable taking the financing aspect of buying an apartment into account.

3.1 Scraping apartment listings

The majority of our data is collected from Bolighed [2018] which is a database for Danish residents on sale containing approximately 67,000 listings. The advantage of Bolighed

[2018] compared to other websites in this class, is that they offer a broad variety of information on each listing such as number of rooms, square metres, energy saving category, how many days the resident has been on sale, etc.

Another advantage of Bolighed [2018] is that they have an API that allows us to easily extract all data from the web page to a .json format. We construct the scraper by iterating over URLs from each page in the search engine and collecting the .json response for every page containing all the necessary information.

After structuring the data, we find that the majority of listings are within the boundaries of the urban area of Copenhagen. Hence, we sort the data set to contain listings with zip codes up to, but not including, 3000, which give us a data set of 3,957 owner-occupied apartments for sale in a dense area spanning 46 kilometres in latitude and 30 kilometres in longitude.

3.2 Ethics of scraping

In terms of ethical considerations, we looked for the robots.txt file of bolighed.dk, but discovered that they do not have any. Therefore, it was up to us to evaluate if there are any ethical issues by scraping their website. We assessed that the data on Bolighed [2018] is publicly available for everyone with internet access and in the listings there are no personally sensitive information. Furthermore, we restrict our search to only gather information on owner-occupied apartments in the area of Copenhagen in order to not scrape too much of their data and to have a more homogeneous sample. Specifically, we only consider apartments with a zip code below 3,000, which results in a sample containing roughly 6 percent of all their data. In the scraping process we include a time limiter, that limits the time rate of our requests to the server to prevent stressing their servers. The code behind the scraper can be examined in Section 1.1 in the Jupyter Notebook file.

3.3 Additional variables

In addition to listing prices on apartments and all their specific attributes, we want to include some variables that can help predict the price in a given location. There are several advanced approaches to do this, but as the goal of this paper is to predict the same

prices as real estate agents, we want to include variables that they might find important. Therefore we include:

- **Energy rating**: Domiciles with a better rating can save money in the long run. Thus, a better rating improve upon the price. Given energy ratings are a letter, we have converted the scale from strings to numeric values, where rating 9 is best and 0 is worst [Franz Fuerst, 2016].[2]

- **Distance to underground**: With Copenhagen's Metro, it is possible to get around the capital relatively fast and the price is quite affordable. Having a domicile close to a station may affect the price positively [McDonald, 2014].

- **Distance to top tier schools**: All parents want their children to attend the best school. Thus, having a top tier school nearby may improve the price. We have scraped a list of all schools in the municipalities we are interested in with a ranking relative to the rest of the schools in its own municipality. From that, we take the top 20 percent schools in each municipality [Donald R. Haurin, 1996].[3,4]

- **Distance to prison**: Living close to a prison may pull the price down, as property owners might feel insecure from the traffic through the neighbourhood by criminals or relatives to the prisoners [Landmark Research, 2018].

- **Floor number**: Living in at ground level usually pull the price a bit down. We find floor number by splitting the address of each listing and look at the last digit.[5]

Besides these variables we also include dummy variables for each municipality to take shifts into account. This can partly be explained by different levels of welfare services or tax rates across municipalities in a ceteris paribus framework [Jonas Zangenberg Hansen, 2018].

It is quite straightforward to find schools, metro stations and jails, but it is more complex to calculate the distances between each apartment and the variable of interest. Fortunately, when we scrape apartment data from Bolighed [2018], we for the most part also retrieve their coordinates. We then use a Python package called GeoPy to retrieve the

[2]See Section 1.2 in the Jupyter Notebook.

[3]The list is from SøndagsAvisen [2015]. The list is three years old, but schools also becomes popular because of their reputation, which they have to build over several years. Thus, it is not of great concern.

[4]See Section 1.3 in the Jupyter Notebook

[5]See Section 1.2 in the Jupyter Notebook

coordinates for the remaining apartments, all top tier schools, metro stations and jail's. From that we calculate the distance between each variable and each apartment. We expect the correlation between distance to top tier schools and metro stations to be negative. As the distance increase, the apartment becomes less attractive. We expect the opposite for jails.[6]

3.4 Total yearly costs of buying

Beside the general features of apartments and their location, we also need to consider the economic aspect. To do this, we consider how buyers finance their purchase as well as the monthly owner expense, i.e. we look at the current expenses of servicing debt and include the owner expense to get the total yearly costs of buying an apartment. To get an estimate of the financing, we will assume that

- The buyer pays 5 percent up front as a cash ticket of the apartments value. The precise amount to pay is rounded to nearest DKK 5,000. This cash ticket is the minimum requirement to pay when buying private real estate in Denmark, see Robinhus [2018].
- The buyer will get a mortgage loan of 80 percent (the maximum loanable mortgage) with 30 years to maturity with an interest rate of 3 percent, see Realkredit Danmark [2018].
- The remaining approximately 15 percent is loaned from a bank with a fixed interest rate of 6.6 percent and 10 years to maturity, see Danske Bank [2018].
- Both loans will be an annuity agreement, i.e. after 30 years all the debt has been paid off.

We add the yearly owner expenses and financing costs to get the yearly total costs for each apartment. In Table 1 in Section 4.1 we have included Total Yearly Costs.

As a digression we can take a glance at the demand side of the housing market by dividing with the average yearly wage before taxes for a representative working family in the urban area of Copenhagen of DKK 832,269 [Statistics Denmark, 2018]. We use a tax rate of 46 percent which has been the average for the period of 2003-2015 [The Danish Ministry of

[6]See Section 1.3 in the Jupyter Notebook.

Taxation, 2017].

$$\text{Spending on housing} = \frac{\text{Mean of Total Yearly Costs}}{\text{Mean of yearly wage after taxes}} = \frac{185,842}{832,269 \cdot 0.54} \approx 0.41$$

This ratio gives an indication of the degree a typical Danish family uses on housing. The rough estimation above shows that a typical representative household in Copenhagen use 41 percent of its income after taxes. This indicates how much a family has for other consumption goods and savings and can also be used as an indicator for how the overall real estate market is priced.

4 Data description

In Section 3 we went through how we obtained data. In this section we examine the distribution and summary statistics of the data. Furthermore, we perform a clustering analysis to group similar observations.

4.1 Descriptive statistics

Completing the data structuring process, we have 3,957 listings of owner-occupied apartments divided over 23 municipalities within the urban area of Copenhagen. Besides information on the specific listing, we also have distance to top tier schools, metro stations and jail's. Table 1 summarises some general statistics of our sample.

The average apartment is priced at DKK 4.1 million, with the most expensive at DKK 25 million and the cheapest at DKK 0.5 million. With an average area of 93 square meter's, it yields an average square meter price of DKK 43,500.

Looking at the cost of living, owner expenses lie in the neighbourhood of DKK 3,700 a month. Comparing the apartment price per square meter with the owner expenses per square meter. The latter has a lower variance, thus, likewise the total yearly costs per square meter has a lower variance. A factor can be that apartment prices reacts more on market movements than the owner expense.

Table 1: Summary Statistics

	Mean	Std. Deviation	Min	Median	Max
Apartment Price	4,164,777	2,569,776	550,000	3,495,000	25,000,000
Owner Expense (monthly)	3,704	1,596	782	3,378	19,756
Total Yearly Costs	185,842	103,082	35,442	15,6904	1,084,716
Sqm. Price	43,467	12,069	10,910	43,013	111,071
Sqm. Owner Expenses (mth)	41	9	10	39	149
Sqm. Total Yearly Costs	1,969	429	564	1,951	4,846
Floor	2	2	0	1	9
Rooms	3	1	0	3	16
Area (m^2)	93	40	24	86	616
Energy Rating	4	2	0	4	9
Days on Market	98	122	0	67	3,536
Distance to top School (m)	1,200	700	0	1,100	5,200
Distance to Metro (m)	3,500	4,700	0	1,800	27,900
Distance to Jail (m)	4,300	4,800	0	2,400	27,800

Data source: Bolighed [2018], SøndagsAvisen [2015], København Metro [2018]

The mean energy rating is 4 (energy-mark C) which we have imputed to the 11 percent of observations with missing variables. The average distance to a top tier school is 1.2 kilometres, 3.5 kilometres to nearby metro station, and 4.3 kilometres to nearby jail.

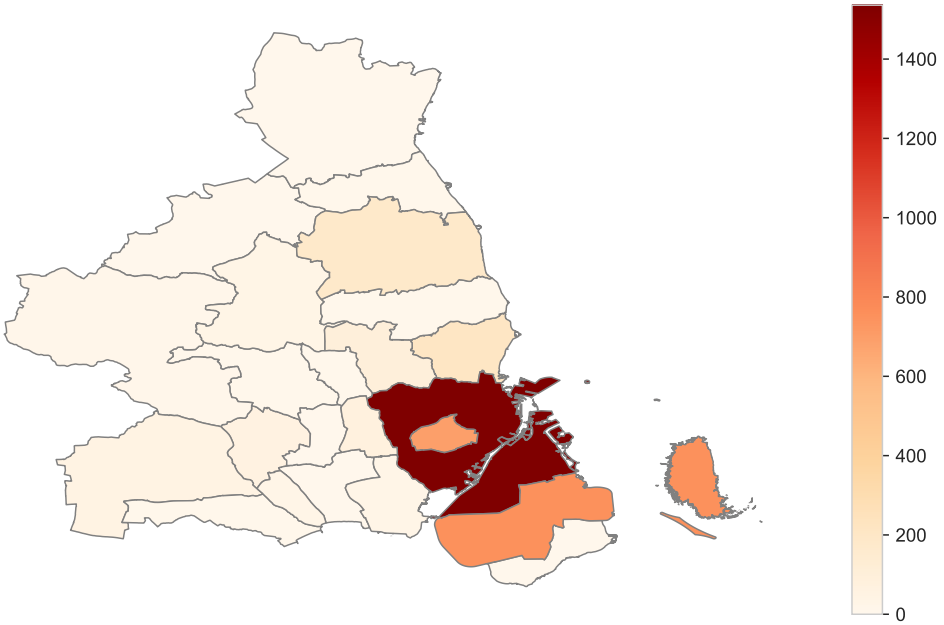
4.2 Distributions of listings and prices across municipalities

Even though we only look at 23 municipalities in the urban area of Copenhagen, there are clear differences in the listings between municipalities. Figure 1 presents a choropleth map of how listings are distributed.[7] Across municipalities the number of listings vary from 3 in Fredensborg to 1,536 in Copenhagen. It is worth noting that the municipality of Copenhagen alone contain 39 percent of all listings. Furthermore, the municipalities of Copenhagen, Frederiksberg, and Tårnby jointly contains 76 percent of the total number of listings.

To get a brief overview of the distribution of square meter prices of apartments across the urban area of Copenhagen, we present the choropleth map in Figure 2. The municipality of Lyngby-Taarbæk yields the highest average square meter price of DKK 58,568. On the other end of the scale, we find the cheapest apartments in Fredensborg with an average square meter price of DKK 14,401.

[7] A choropleth map is a thematic map, where areas are shaded in proportion to a statistical measure.

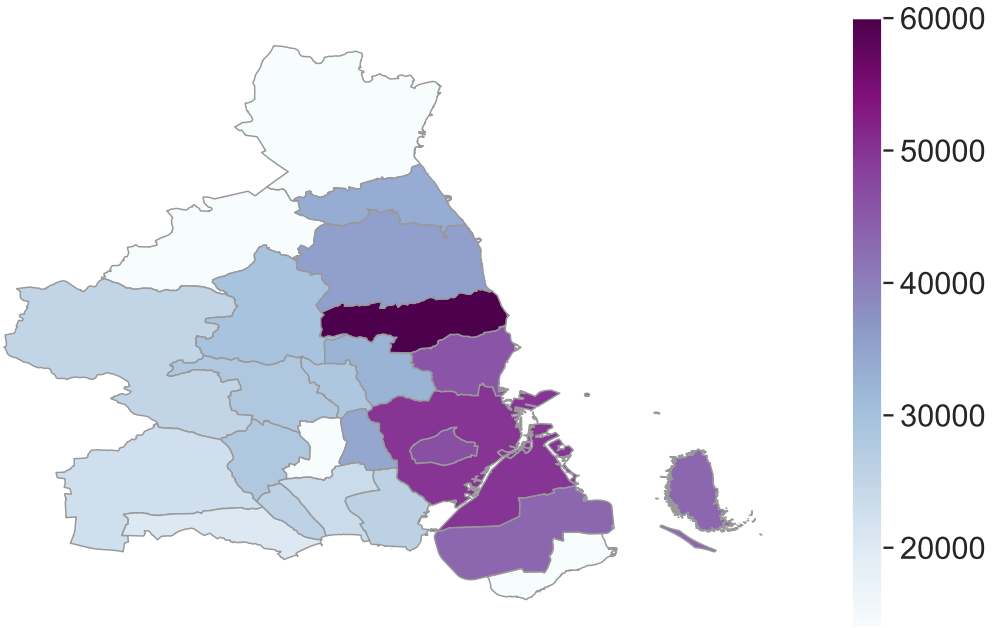
Figure 1: Distribution of listings across the urban area of Copenhagen



Data source: Bolighed [2018]

Generally, the most expensive apartments are close to the centre of Copenhagen. In the municipalities of Copenhagen, Frederiksberg, Gentofte and Tårnby, the average square meter price is around DKK 45,000.

Figure 2: Distribution of square meter price across the urban area of Copenhagen



Data source: Bolighed [2018]

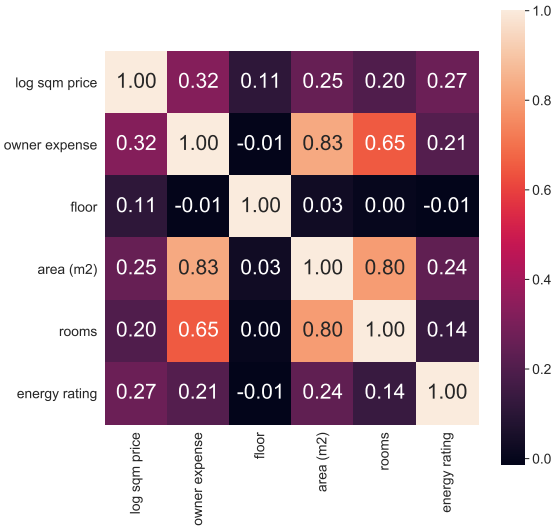
4.3 Correlations

The goal of this paper is to predict the listing prices. For this purpose, we transform our target value, *square meter price*, into a logarithm as a remedy to positive skewness.

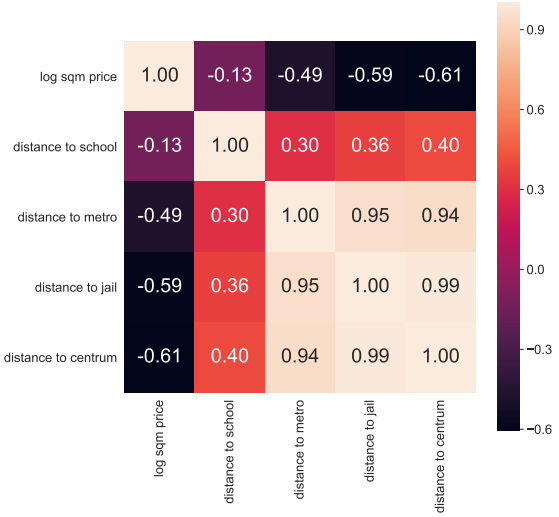
To examine how the features are linearly correlated with the target, we create two heat maps presented in Figure 3, using the NumPy function .corrcoef in Python. Most of the classic apartment properties from Panel (a) have a weak positive correlation with log square meter price, the strongest correlation being 0.32 with owner expenses.

Figure 3: Correlation matrices of target and feature variables

(a) Apartment properties



(b) Distance measures



Looking at the distance measures in Panel (b), all are negatively correlated with log square meter price and the magnitude of the correlations are higher. For metro stations, top tier schools and centre this is in accordance with our expectations; the listing price should be lower when the distance increases to a top school, good transportation, or simply the centre. However, the jail variable is also negatively correlated, which is surprising. This might be explained by the fact that prisons are located quite central as the matrix shows distance to jail and distance to centre are almost perfectly correlated. So, in a multiple regression analysis, where we control for other features, the local effect of living close to a prison might still decrease the listing price.

4.4 Clustering of coordinates and prices

Having examined the distribution, mean and correlation of the data set we want to further explore the listing prices, which we want to predict, using *clustering*. Specifically, we will examine the effect of location. We expect location has a significant effect on house prices.

Clustering is a technique that allows us to find groups of similar objects, objects that are more related to each other than to objects in other groups. We use clustering to visualise how strong the relation is between the locations and the square meter prices of apartments. Furthermore, clustering gives us the coordinates of the centroid (cluster center) of the high-price cluster, from which we construct a new variable *distance to center* as a proxy for the distance to an expensive location.

4.4.1 K-means clustering

We use the clustering algorithm K-means. The K-means algorithm belongs to a category of prototype-based clustering. Prototype-based clustering means that each cluster is represented by a prototype, which can be the centroid of similar points with continuous features, or the medoid in the case of categorical features. While K-means does well at identifying cluster's with a spherical shape, a potential drawback of this clustering algorithm is that we have to specify the number of clusters, K. We have done that by running the algorithm with different numbers of K and looked at which amount gives the best visual result. To calculate which objects has the same similarity, we use the squared Euclidean distance between two points.

$$d(x, y)^2 = \sum_{j=1}^{m} (x_j - y_j)^2 = |x - y|_2^2$$

Based on the Euclidean distance, we can describe the K-means algorithm as a simple optimisation problem, an iterative approach for minimising the within-cluster sum of squared errors, which is sometimes also called cluster inertia:

$$\mathbf{SSE} = \sum_{i=1}^{n} \sum_{j=1}^{k} w^{(i,j)} |x^{(i)} - \mu^{(j)}|_2^2$$

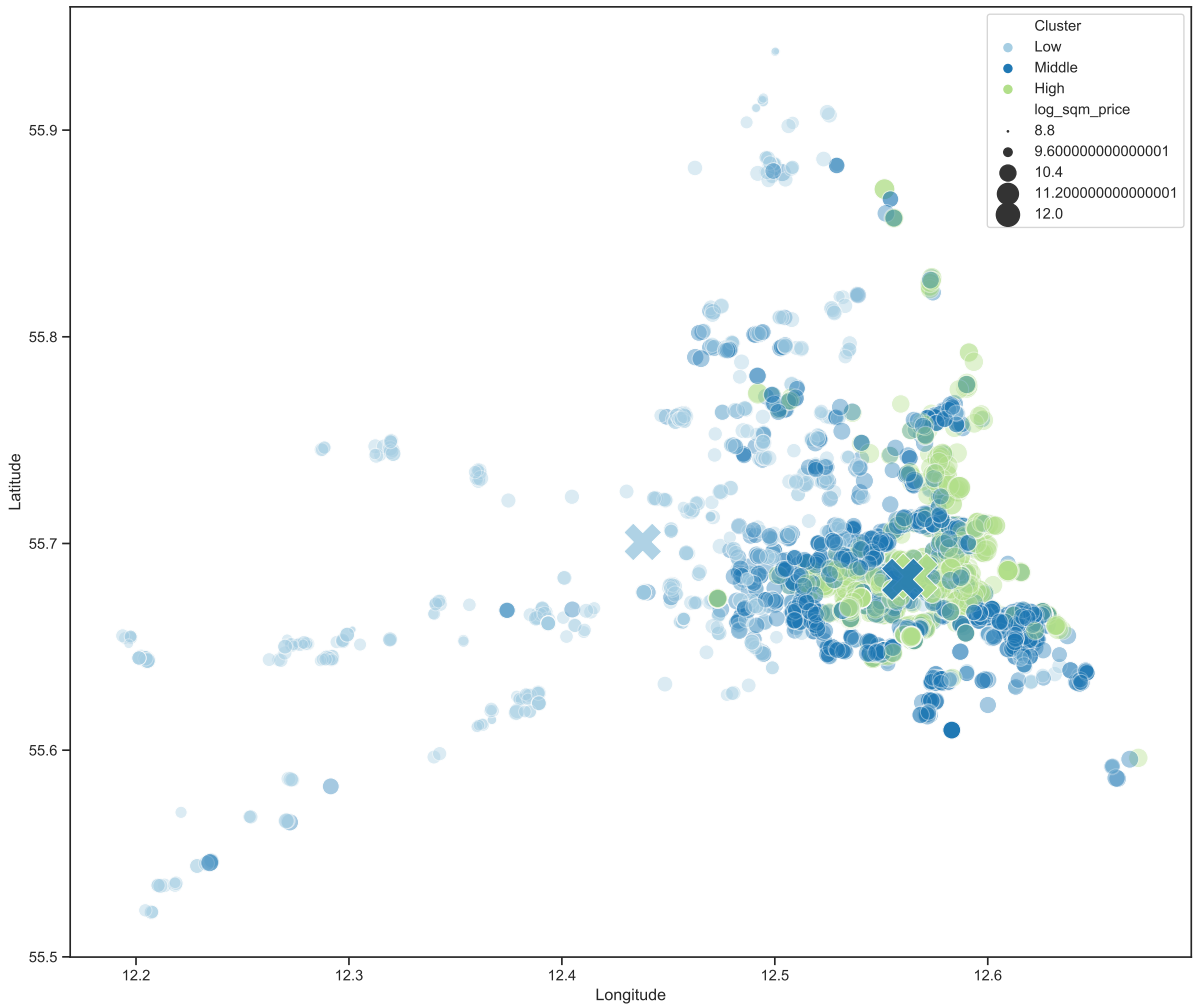
4.4.2 Clusters in the urban area of Copenhagen

While K-means clustering can be performed in any N-dimensional space of numerical variables we choose to perform it for a three dimensional space to examine the relation between latitude, longitude, and logarithm of the square meter price of apartments, without shifting the weight to other features. This allows us to examine the clusters of low-priced, middle-priced and high-priced apartments. In Figure 4, we have plotted the centroids (marked by X) and apartments demarcated by their cluster as a scatter plot with latitude and longitude on the axis. The algorithm runs in less than a second and for the seed numbers from 0-500 the number of iterations before convergence is ranging from 5 to 31.

The centroid of the high-price cluster is located exactly at Nansensgade 53 in Inner City. The centroid of the low-price cluster is far west of the Inner City and the eastern coastline, though a few low-price apartments are also placed to the south east on the island of Amager. The middle-price cluster is less than 400 metres to the south west of the high-priced centroid, though they have a larger distance to the centroid than high-priced cluster, and is placed more or less as a ring around the centre.

It does not provide meaningful information to increase the number of cluster's, because the new centroids will lie very close to the former, which will blur the patterns of the cluster members. This shows that there is a limit to cluster analysis when the relation between variables is nonlinear and complex. Furthermore, the fact that two of the centroids are in the centre of the data set is also biased as the method is volatile to there simply being more observations there while the average price is higher further north in Lyngby-Taarbæk as seen in figure 3.

Figure 4: Clustering of coordinates and log square meter prices



Data source: Bolighed [2018]

Note: Apartments for sale demarcated in three clusters given latitude, longitude, and the natural logarithm of the square meter price of owner-occupied apartments. The cluster centres are marked by X. The dot size increases with the log square meter price as an indication of the third dimension of the data space.

While the clusters are slightly blurry when their coordinates are mapped (Figure 4), we see from Table 2 that omitting the 5 percent most extreme observations in each cluster there is no overlap between the prices levels of square meter price.

Table 2: Distribution of square meter price for each cluster

Cluster	Count	Mean	Std	Min	2.5 pct.	50 pct.	97.5 pct.	Max
Low-price cluster	813	28,130	4,020	10,910	18,060	28,951	33,272	34,322
Middle-price cluster	1,728	40,480	4,035	31,933	33,544	40,690	46,893	47,204
High-price cluster	1,416	55,916	8,773	47,181	47,500	53,429	80,037	111,071

From Table 2, we see that 44 percent of the apartments are allocated to the middle-price cluster, 36 percent are allocated to the high-price cluster, and 20 percent to the low-price cluster.

As a robustness test we also tried assigning clusters based on the log total yearly expenses per square meter instead, but as we know from Table 1, the variation in owner expenses per square meter is smaller than for the apartment price per square meter. Thus, the middle-price cluster will make up 48 percent of all apartments. We choose not to use this clustering as we prefer a more evenly spread cluster. In total 11.4 percent of the apartments are allocated to a different cluster, which shows that the clustering is quite robust.

Summarising, we see that location seem to have significant effect on prices per square meter and has to be taken into consideration when predicting.

5 Machine Learning methods

In the previous section we went through our data set. In this section, we outline the methods we use to predict listing prices of owner-occupied apartments. We use a machine learning approach and compare the results of the linear regression with OLS.

The goal is to predict the listing prices on the apartments on sale. Our target variable is *square meter price*, with a range of features including *number of rooms*, *floor*, *area*, *garden area*, *owner expense*, *energy rating*, dummy variables for *municipality*, *days on market*, and measures of distance to *top tier schools*, *metro stations*, *prisons* and distance to *centre*. Our approach is to first train a linear regression model on our data. Next, we introduce a model with Lasso regularisation as a remedy to over-fitting. This is especially important when we introduce polynomial features in the regression as the number of interaction-terms increases exponentially. The polynomialfeatures package not only creates a polynomial of a specified order for each variable but also creates pairwise interaction terms for all variables. Next, we apply K-fold cross validation, to see if we can optimise over hyper-parameters and whether providing more training sets results in a more robust model. Finally, we compare which of the models perform best at predicting the listing prices of owner-occupied apartments. We will accomplish this using a number

of packages from the sklearn module in Python.[8] The accuracy of our models will be evaluated by the **Mean Actual Error** (MAE) and **Root Mean Squared Error** (RMSE) for each approach:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \hat{y}_i\Big)^2}. \tag{5.2}$$

MAE measures the average absolute size of the prediction errors without considering their direction. RMSE is the square root of the mean squared errors, such that the magnitude of the errors are taken into account, as we square the errors before taking the mean. In comparison, this means that RMSE attaches higher weight to large errors [Chai, 2014]. For both measures we interpret a smaller number as a more precise predictive model.

5.1 Linear Regression (OLS)

OLS minimise the sum of squared errors and can be useful for finding in-sample causal relationships when properly controlling for biases arising from unobservables and self-selection [Angrist et al., 2013]. When building a model with the purpose of a strong prediction power, however, we have to be careful to balance variance and bias. If a model performs well in-sample but has poor prediction power, it is likely over-fitted. This is a consequence of having too many parameters which causes high variance and makes it oversensitive to spurious patterns. On the other hand, we have to be careful that the model becomes too simple and suffer from under-fitting. This is the case when our model is not complex enough and highly biased, which results in poor prediction out-of-sample as well as it does not find all relevant patterns in the data. On out-of-sample data, OLS does not generally perform very well. We expect that our linear regression will most likely be over-fitted due to too many interaction terms.

[8] The machine learning techniques can be examined in Section 3 in the Jupyter Notebook.

5.2 Lasso Regularisation

Regularisation help us tackle the problem of over-fitting by either reducing the number or size of coefficients, which is achieved by including a penalty term against the complexity in the optimisation process,

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}_{\text{MSE}} + \underbrace{\lambda \sum_{j=1}^{m} |w_j|}_{\text{Penalty}}. \tag{5.3}$$

The Lasso approach penalises against complexity by setting some of the weights to zero if λ is high. λ is the hyper-parameter which determine to what degree we should penalise OLS [Raschka and Mirjalili, 2017]. We use two variations of the Lasso model: First using the default value of $\lambda = 1.0$ in a simple model, and next use cross validation to optimise λ, which hopefully results in an improved model.

5.2.1 K-Fold Cross Validation

We use K-fold cross validation to optimise over hyper-parameters and find an appropriate penalty rate, λ, and the amount of polynomial features.

The K-fold method is performed by splitting our data into a development set and a test set, where the test size is set to 1/3. The development set is then split into K random equal sized folds. We use $K - 1$ folds for training and 1 fold for validation, and then rotate the part of the data that is used to train and validate the model. This way, our model choice will not depend on which sample we initially pick. Further, providing more training samples to the learning algorithm usually result in a more robust model. We will use $K = 10$ folds as suggested by Ron Kohavi [Raschka and Mirjalili, 2017].

6 Results

In the previous section we went through the theory of the methods we use to predict listings prices. In this section, we present our main results for the different models and examine whether we do well at predicting the same square meter prices as professional real estate agents, making them redundant.

6.1 Prediction results

The first results are presented in Tabel 3, where we choose a constant polynomialfeature of order 3. The MAE and RMSE are interpreted as an approximate percentage deviation due to the logarithmic transformation of the price per square meter.

Table 3: Prediction of log price per square meter with 3 polynomials

		MAE	RMSE	λ	Polynomial
(1)	Linear	$8.10 \cdot 10^{14}$	$2.88 \cdot 10^{16}$	-	3
(2)	Lasso	0.222	0.282	0.100	3
(3)	Lasso CV	0.123	0.164	0.003	3

Note: The results can be interpreted as an approximate percentage deviation from the true listed square meter price.

In accordance with our expectations, the linear model in (1) performed poorly. Both MAE and RMSE explodes and have no meaningful interpretation, because numbers far from zero no longer be interpreted as percentage deviations. This suggests that the model is over-fitted and have poor out-of-sample prediction power.

The Lasso model in (2) did better than the linear model without optimising over λ. We get a MAE of 0.22. This corresponds to a prediction error of approximately 22 percent, which is a major improvement from the linear model, but still not good.

Our best prediction comes from the Lasso CV model in (3), where we use K-fold cross validation to optimise over λ. A MAE of 0.122 corresponds to a prediction error of 12.3 percent. In our sample, for an average apartment of 93 square meters with a square meter price of DKK 43,500, a 12 percent deviation accumulates to approximately an error of DKK 485,000 per listing. This is a significant error, and greatly exceeds the potential savings by excluding the real estate agent.

6.2 Model validation

The accuracy of the linear model is horrific and might be explained by the number of polynomials we include. Instead, we train the model again, but this time while optimising over polynomialfeatures of order 1 through 5 and picking the model with the lowest RMSE value. The results are shown in Figure 4.

Table 4: Predicting log price per square meter optimising over polynomials

		MAE	RMSE	λ	Polynomial
(4)	Linear	0.136	0.182	-	1
(5)	Lasso	0.222	0.282	0.100	1
(6)	Lasso CV	0.123	0.164	0.003	3

Note: The results can be interpreted as an approximate percentage deviation from the true listed square meter price.

The linear model in (4) performs much better with a first degree polynomial compared to (1), and even performs better than the simple Lasso model. This is because model (4) is no longer subject to over-fitting, so the Lasso has nothing to penalise. However, model (4) is not better than (3). This might be because it is too simple and does not capture enough of the nonlinear relationship in the data set.

Model (6) in Tabel 4 corresponds to our optimal model (3) in Table 3. The cross validation chooses three degrees of polynomialfeatures, which is exactly what we manually chose before. This Lasso CV is our preferred model with a $\lambda = 0.003$ and polynomial of order 3, resulting in a MAE of 0.122.

We get a MAE of 12 percent, thus, if we were to value an apartment on using our model, it would result in an error just below DKK 0.5 million on average, much more than the cost of a real estate agent (of around DKK 100,000).

Even our best model lacks accuracy, but a machine learning approach has the potential to assist either real estate agents or households when valuating residential prices. Though, the model should be more well-specified and probably include more variables. This will further be discussed in Section 7.

7 Discussion

In our results presented in the previous section we saw that our model has an error of DKK 0.5 million for an average apartment. In this section, we will discuss our results, and how we might improve upon our model.

7.1 Validity of predictions

Our model predicts an error of DKK 0.5 million for an average apartment in our sample. If we assume real estate agents' valuations are equal to the true prices, an error of DKK 0.5 million greatly exceeds the potential savings of handling the sell yourself as one have a high risk of either setting it below the true value or above the demand price. However, real estate agents are also prone to errors. Thus, our prediction can be even further from the true market value or closer to it. A proxy for real estate agents error is the difference between the listing price and realised price. According to Finans Danmark, this difference was around 0.1 million in 2017.[9,10] Our model slightly under-predict the real estate listing price (on average). As real estate agents generally over-predict, the model comes closer to the true price than it does to the targeted listing price.

Nonetheless, our model still does not have satisfyingly precise predictions, and we need a more well-specified model. The root mean squared error (RMSE) suggests that the models are having a hard time predicting some of the apartment's unique specifications. This issue might be improved by including more features such as being located near green areas, or next to busy main roads[Per Thiemann, 2014]. However, the MAE and RMSE will always have at least the same magnitude and the difference between these measures in our models are not immense.

Instead, we might focus on extending the model selection process to see if there exists a better model in the class of supervised learning models, that more accurately predicts the listing prices. Some examples are Ridge regularisation or Random Forest.

Another aspect of the validity of our model is the direction of the error. We would hope that when we predict a higher listing price than the actual, it is just poor luck, but when we predict a lower listing price than the actual, the average days on market should hopefully be greater than when we over-predict. This would mean our model might be able to predict when real estate agents set the listing price too high. A too high listing price can make it more difficult to sell an apartment. Unfortunately, our best prediction model yields an average days on market of 99 for errors above the true value and 88 days for errors below the true value. Thus, our model does not do well at capturing

[9]The difference is between first listing price and realised price. Thus, it is not the true error, because we do not know whether the apartment was listed by an authorised real estate agent our by self sale.

[10]See Table BM010.

overestimated prices.

7.2 Possible extensions

The difference between the MAE and the RMSE values indicate that we might have an issue with predicting some of the extremes. Excluding the bottom 25 and top 25 square meter prices from the sample decreases the prediction error (MAE) from 12,2 percent to 12,0 percent for the Lasso CV with three degrees of polynomial features (Appendix, Table I.1), however we would also like our model to be able to predict the extremes.

Some natural extensions to our models would be to include more apartment specific attributes such as whether the apartments have balconies, the conditions of the apartments, or the age of the building. The latter information is publicly available while data on balconies and conditions are not. These and examples of the information that real estate agents collect. As a proxy for the general conditions and non-quantifiable features we could look at the level of prior realised prices[11]. We would then have to control for level shifting in the market over time.

A simple general way to take more location attributes into account as those discussed in Section 7.1 is to directly include latitude and longitude as variables as we might expect the price to increase with longitude and with the interaction of the two (i.e. a higher price to the east and north east). This gives an improved MAE of 12,1 percent for the Lasso CV with three degrees of polynomial features (Appendix, Table II.2).

Another interesting extension, which is related to discussion in Section 7.1, is to investigate how precise real estate agent's calculate the market value of the house, i.e. how far they are from the true price. To explore this we would need to scrape the realised prices instead.

To put it in a more macro context another extension would be to investigate how the overall real estate market is priced. Is it priced too high or can the prices keep increasing as they have done in Copenhagen City since mid 2011? This could be investigated by looking at the degree of how much they use on housing of their yearly income, as we looked at in Section 3.4. Macroeconomic variables like economic growth, unemployment

[11]Like for the age of the buildings information on prior sales are only available for apartments that have been traded since 1992, e.g. at ff

rate, purchasing power etc. could potentially help to indicate shifts in the demand curve over time.

8 Conclusion

The goal of this paper is to predict listing prices of owner-occupied apartments in the urban area of Copenhagen. Clustering of prices and coordinates as well as mapping of average prices in municipalities visualise that location highly affects the prices. We scrape apartment and location data data from different internet sources, transform and clean it. We then apply different machine learning methods and test each model taking a simple OLS regression as a benchmark.

Our best model for predicting listing prices is a Lasso CV model where we use a K-fold cross validation to optimise over λ and polynomial features. We get a prediction error of 12.3 percent which for an average listed apartment corresponds to an error of DKK 0.5 million from the target. As the saving potential is around DKK 100,000, and the average error of real estate agents is in range of DKK 100,000 our model is not precise enough to leave out traditional valuation methods and discard real estate agents.

However, the methods we have been testing could be calibrated to be more precise with more data and different machine learning methods.

9 References

Angrist, J. D.; Pischke, J.-S., and Pischke, J.-S. *Mostly harmless econometrics: an empiricists companion*. Cram101 Publishing, 2013.

Bolighed, Retrieved August 2018. URL https://bolighed.dk.

Bolius, Retrieved August 2018. URL https://www.bolius.dk/saa-meget-koster-det-at-saelge-sit-hus-8664/.

T.Chai, D. P. Root mean square error (rmse) or mean absolute error (mae)? 2014.

Danske Bank, Retrieved August 2018. URL https://www.danskebank.dk/PDF/Priseksempler/Priseksempel-Boliglaan.pdf.

Donald R. Haurin, D. B. School Quality and Real House Prices: Inter- and Intrametropolitan Effects. December 1996.

Ester Hansen, M. H. R. and Staghøj, J. Boligboblen der bristede: Kan boligpriserne forklares? Og kan deres udsving dæmpes? *Kvartalsoversigt - 1. kvartal 2011 - Del 1*, 2013.

Pat McAllisterFranz Fuerst, A. N. P. W. Energy performance ratings and house prices in Wales: An empirical study. May 2016.

Gattini, L. and Hiebert, P. Forecasting and assessing euro area house prices through the lens of key fundamentals. 2010.

Andreas Østergaard IversenJonas Zangenberg Hansen, P. S. Real estate in the 21st century. 2018.

København Metro, Retrieved August 2018. URL https://www.m.dk/#!/.

Landmark Research, Retrieved August 2018. URL http://digital.library.wisc.edu/1711.dl/RealEstate.LRTrostel.

Masías, V. H.; Valle, M.; Crespo, F.; Crespo, R.; Vargas Schüler, A., and Laengle, S. Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile. January 2016.

Matthew Cooke, Retrieved August 2016. URL https://www.telegraph.co.uk/property/online-estate-agent/how-much-is-my-house-worth/.

McDonald, D. P. M. J. Reaction of House Prices to a New Rapid Transit Line: Chicago's Midway Line, 1983–1999. August 2014.

Per Thiemann, Retrieved August 2014. URL https://politiken.dk/oekonomi/bolig/art5502020/M%C3%A6gler-Her-er-faktorerne-der-tr%C3%

A6kker-din-bolig-mest-op-i-pris.

Raschka, S. and Mirjalili, V. *Python Machine Learning*. Packt Publishing, 2017.

Realkredit Danmark, Retrieved August 2018. URL https://www.rd.dk/da-dk/privat/koeb-bolig/Kurser-og-renter/Pages/Aktuelle-priser.aspx?segment=P&obl=fl.

Robinhus, Retrieved August 2018. URL https://www.robinhus.dk/fakta/finansiering/brutto_netto.asp.

Lynn MacDonaldStacy Sirmans, D. A. M. A meta-analysis of selling price and time-on-the-market. *Journal of Housing Research*, 19(2):139–152, 2010.

Statistics Denmark, Retrieved August 2018. URL https://www.statistikbanken.dk/statbank5a/SelectVarVal/saveselections.asp.

SøndagsAvisen, Retrieved August 2018 2015. URL https://www.sondagsavisen.dk/familien/2015-08-22-se-hele-listen-her-er-danmarks-bedste-og-vaerste-skole/.

The Danish Ministry of Taxation, Retrieved August 2017. URL @misc{agent_val,author={{MatthewCooke}},year={2016},month={RetrievedAugust},url={http://www.skm.dk/skattetal/statistik/generel-skattestatistik/skattetrykket-i-eu-landene}}.

Appendices

I Prediction excluding extremes

Table I.1: Predicting price per square meter excluding the 50 extremes

		MAE	RMSE	λ	Polynomial
(1)	Linear	0.133	0.169	-	1
(2)	Lasso	0.210	0.259	0.100	1
(3)	Lasso CV	0.120	0.155	0.003	3

II Prediction including latitude and longitude

Table II.2: Predicting price per square meter with three polynomials

		MAE	RMSE	λ	Polynomial
(1)	Linear	0.128	0.172	-	1
(2)	Lasso	0.222	0.282	0.100	1
(3)	Lasso CV	0.121	0.161	0.003	3