DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Introduction to Social Data Science exam paper

Price Valuation of Leisure Properties in Denmark

Exam numbers

78, 191, 208, 217

Contributions

191: 1.1, 2.3, 2.4, 4.4, 5.2 , 5.3 , 6
78: 1.1, 3.1, 3.2 , 3.3 , 3.4 , 3.5 , 6
217: 2.1, 2.2.1, 2.2.2, 2.2.4, 4.4 , 6
208: 1.2 , 2.2.3, 4.1, 4.2, 4.3 , 5.1, 5.4 , 6

character count: 38076

August 23, 2021

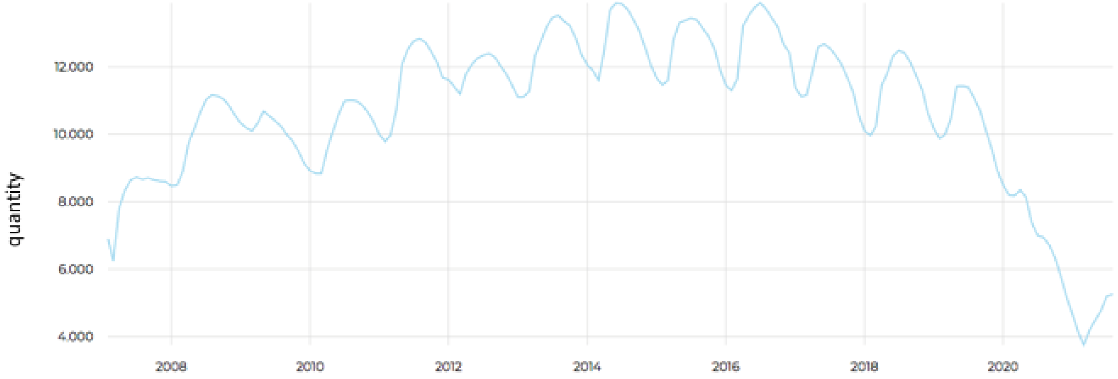
Contents

1 Introduction

1.1 Problem motivation

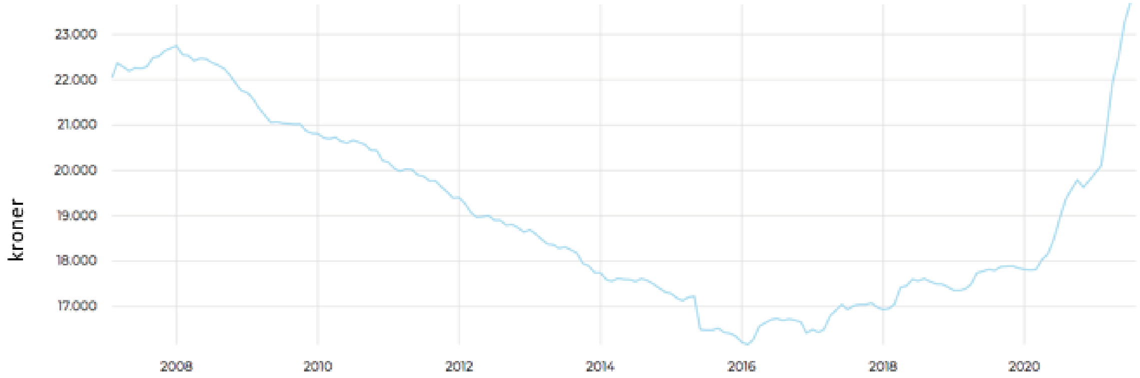
Since the financial crisis, the market for leisure properties has gone through major changes. It is clear that from 2008 the number of listed leisure properties has gone up, while the average price per square meter has gone down (see figure 1 and figure 2).

Figure 1: Amount of listed leisure properties



Source: Boliga.dk.

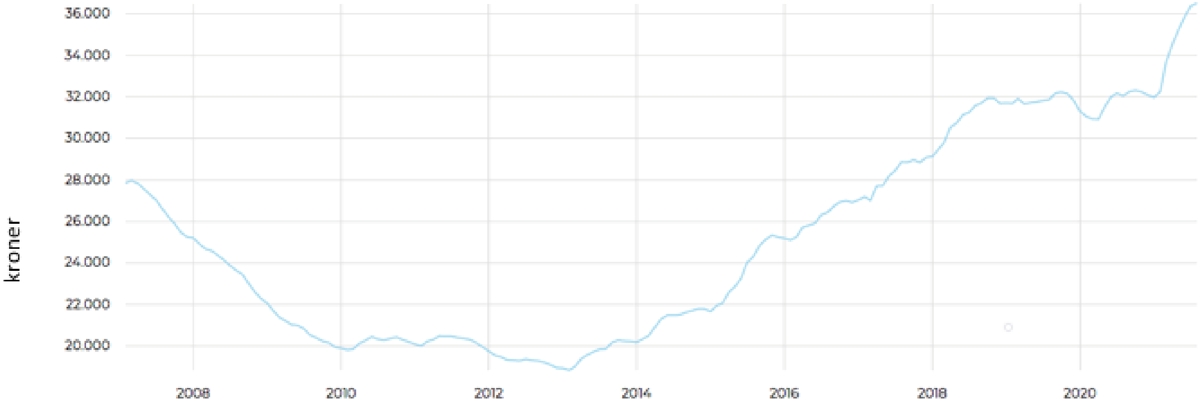
Figure 2: Average price per square meter, leisure properties



Source: Boliga.dk.

This development seems very similar to the general housing market. However, the leisure property market differs from the general housing market in its recovery following the financial crisis. While the average square meter price on apartments reaches its minimum in 2013 (as seen in figure 3), it is not until 2016 that the average price of leisure properties stops decreasing (see figure 2). The price of leisure properties increases with a steady pace from its minimum in 2016 until the corona lockdown in 2020, which supposedly increased the demand of leisure properties, rapidly driving the price up.
Furthermore, by comparing the development in square meter prices, we observe that the general housing marked has superseded the peak level from before the financial crisis by roughly 20 percent, whilst leisure properties have superseded their previous all time high by a smaller margin of roughly 5 percent.
The development indicates that the demand for leisure properties has been low following the financial crisis of 08'. However, it would seem that the corona crisis has driven up demand and prices. Furthermore, the volume of sales has increased since the beginning of lockdown. This development of higher prices and increased volume of sales means that the market for leisure properties stands for an increased amount

Figure 3: Average per square meter, apartments



Source: Boliga.dk.

of economic transactions within the last year. The increased sales volume also makes leisure properties more liquid as an investment, presumably driving demand up due to low interest rates.
With this in mind it is interesting to construct a tool to efficiently value leisure properties based on location and other geographical data. This leads us to the *research question:*

Can geographical features be used to accurately predict real estate price for leisure properties in Denmark.

1.2 Literature Review: Big Data & Machine Learning

The continued growth in technology the past decades has resulted in a large amount of data being generated at a rapid pace. The advances that has been made on mobile devices, digital sensors, communications, computing, and storage has aided the process of collecting data. This is one of the key motivations for current and future research (Yagoob, Ibrar et. al., 2016). According to Doug Laney, he defined in 2001, Big Data by the "3 Vs": *volume, velocity* and *variety.* Volume describes the size of the data set while velocity indicates the pace of data going in and out. Lastly, variety describes the diversity of data types and sources. Under special conditions the 3 Vs are extended to include *variability.* This refers to how cluttered and reliable the data is (Chen & Zhang, 2014).

Big Data can be used to create models that predict certain values of interest. One popular use is predicting house prices based on several socio-demographics characteristics. The literature suggests that the trending method when estimating house prices is through a hedonic-based regression or artificial intelligence techniques to predict the price. Several studies suggests that the hedonic price model has been used to identify the relationship between house prices and its characteristics. However, the potential limitation of this method lies in the fundamental model assumption and its estimations. These assumptions are identification of supply and demand, market segmentation, market disequilibrium and selection of independent variables (Park & Bae, 2015).

Recent studies suggests that using machine learning algorithms to predict the housing market may be a potential competitor. By using a neural network technique the model is capable of recognizing statistical patterns when examining the relationship between house prices and characteristics (Fan et al., 2006). In some cases the neural network prediction model performs better compared to the hedonic price model when predicting real estate prices (Selim, 2009). Aside from the neural network technique there are several other machine learning algorithms such as linear regression, least absolute shrinkage and selection operator (LASSO) and gradient boosting. Satish et al. (2019) showed that LASSO regression performs better compared to the other algorithms in terms of predicting house prices. Compared to the conventional econometrics models, machine learning algorithms do not require the training data set to be normally distributed. In general, the statistical tests used in conventional econometrics models depend on the assumption of a normal distribution. Without this assumption the tests and results are not viable (Winky et al., 2021).

Overall, machine learning algorithms give rise to several benefits in research as they can provide a more flexible and sophisticated estimation approach when analyzing data with a large amount of variables. Large data sets tend to be challenging when estimating with classical a linear model, such as ordinary least square (OLS), and may provide subpar results notably due to overfitting.

2 Data Collecting

2.1 Data Sources

A list of all the variables scraped and used, as well as a brief description is presented in the appendix (see table 4). We collected data from three main sources:

2.1.1 Boliga.dk / dingeo.dk

Boliga.dk is one of the largest websites for real estate sales in Denmark, which we favoured over the alternative boligsiden.dk. We intended to use some variables from the website dingeo.dk (owned by Boliga), which lists an enormous amount of information on individual addresses in Denmark. However, the sheer amount of available information and the specific layout of the website made it difficult to scrape. Instead, we opted to scrape the relevant information directly from the individual boliga.dk adverts. This makes us indirectly use dingeo.dk's API, which boliga.dk uses to get the data it displays. Collected data based on the sales adverts include; address, zip code, property and house size, boliga ID, price, etc. The extra data from dingeo.dk examines risks for health issues, flooding, and crime for the individual address.

2.1.2 Hvorlangterder.dk

This website was created by the startup Viamap, specialising in map-based algorithms for private firms. Given an address, it returns the distance to 20 points of interest to homeowners: nearest grocery store, bus stop, forest, shore, hospital, school, football field, etc.

Looking at a map of leisure homes in Denmark, it is clear that a large part of them are located outside the big cities, and many are near a body of water. This indicates that the real estate adage "Location, location, location" holds true for leisure homes in Denmark. As such, we hypothesise that the distance to these points of interest can serve as significant variables in predicting the valuation of the house.

In our scraping process, we will input all the addresses collected from our scraping of boliga.dk into hvorlangterder, and collect the returned data to add it to our dataset.

2.1.3 DMI Weather archive

The Danish Meteorological Institute hosts an online weather archive with historical data on weather patterns: hours of sun and rain, wind direction and speed, location of lightning strikes, pressure, and temperature. This data is extremely high-frequency, measured down to the level of the hour for the past 10 years. Unfortunately for us, it is only available at the level of the municipality, rather than at that of the individual town.

Danish summers are infamous for their fickle weather, and we wish to investigate whether weather patterns can actually be used to better predict prices for leisure properties in a given area. This in particular, we believe, sets our model apart from others in the field. We may also find that it is in fact a completely irrelevant variable for the valuation.

2.1.4 Additional data sources

- **postnord.dk**: From the Scandinavian Post Service, we used a dataset matching Danish zip codes with their municipality. This data was not a part of the boliga adverts, but was necessary to link them with the DMI data, only available at the municipality level.

https://www.postnord.dk/kundeservice/kundeservice-erhverv/om-postnumre/postnummerkort -postnummerfiler

- **dingeo.dk**: This website, owned by the Boliga company, hosts the aforementioned risk data regarding health, crime, and natural disasters for individual addresses. It is then reused by Boliga on their website, which we scrape.

2.2 Data Collection Process

Our data collection process can be divided into 4 parts. *Boliga search results, Boliga individual adverts, Hvorlangterder* & *DMI weather archive.* [1].

2.2.1 Boliga search results

The first step is to get a list of all the individual leisure homes available for sale. To do so, we search for all adverts on boliga.dk, using their filters to return only leisure homes. This search was conducted on August 20th 2021, and returned **5283** adverts all over Denmark.

The search page communicates with boliga's internal API, and the response includes **18** useful variables in the metadata, including: Location data, size, down payment, specific physical characteristics of the house, and IDs for Boliga's database as well as the GPS system DAWA. From this, we also got our target variable: The listed sale price.
For this reason, it is unnecessary to scrape the individual pages to get most of the wanted information, which saves massively on scraping time. Exactly what variables are scraped and where is discussed below.

2.2.2 Boliga individual adverts

From the Boliga search results mentioned in the above section each individual leisure house's page was scraped. Each returned an URL from the search and was scraped for the data regarding health (tied to pollution), crime, and natural risks. This yielded a total of **9** new variables. This is data that Boliga advertises as new on their website, and that they source from the website dingeo.dk. We chose to scrape this information from Boliga rather than from dingeo.dk directly, as the latter's pages contained far more irrelevant information which also happened to be formatted in a way that did not favour scraping. Scraping Boliga directly proved easier, faster, and less tedious.

2.2.3 Hvorlangterder

Hvorlangterder.dk returns distances, based on the address, to everyday commodities such as schools, supermarkets and hospitals. The scraping was based on a function which takes the obtained address from Boliga's search results and enters it in the search bar. The function returned **20** variables where relevant information was extracted for further analysis.

This website did unfortunately not contain an API, or at least not one we could easily find in the source code. Instead, we used a webdriver to manually go through each search. This took a longer time (2.5 seconds / search), but we did not encounter any errors with this process.

[1] Further details regarding the exact way we scraped the individual pages can be found in the relevant section in the Jupyter notebook containing all our code turned in alongside this report.

2.2.4 DMI weather archive

As stated earlier, we collected historical weather data at the level of the municipality dating back to 2011. We used the website's API to collect the data on temperatures, hours of sun, precipitations, humidity, and pressure.

As our analysis is not concerned with historical trends, we chose to aggregate the data for each municipality. We took the mean yearly value of each of the **7** variables, and used this in our model.

2.3 Data Cleaning and transformation

The scraping process generated a data set with a total of5283 observations. However, some data cleaning was needed before conducting the analysis. The process mostly consisted of converting string to integers, handling missing values and dropping irrelevant parameters. The following section will go into more detail of how this was achieved, along with the thought process behind them. Due the the specification of leisure homes, and not the entire real estate market our retrieved data set contains a relatively small amount of observations. In general our mindset has been to save as many observations as possible due to the size of the data set.

2.3.1 Initial clean-up

The initial cleanup section will cover our correction of errors in the data generating process, such as missing values due to issues with interactions between the different data sources.

At the initial stage of the clean-up process the first step was to make an attempt to fix data that did not return values from Hvorlangerder.dk.
The issue was that some of the collected addresses had either some error in the listed address, no address at all or a format not suited for the search function on Hvorlangterder.dk. 940 listings did not return results from hvorlangterder.dk, and for a dataset of 5283 observations this would be a significant chunk of data to throw away if we were to discard the observations with missing data. For that reason we salvaged as many observations as possible using string operations to construct new addresses compatible with hvorlangterder.dk. We managed to successfully correct roughly 780 listings leaving 160 observations which we would later drop. Some of the later dropped observations included very recently built or project stage properties which hvorlangterder.dk could not retrieve information on at this point of time.

Furthermore, some retrieved data from boliga.dk had errors in their construction year' variable. Some observation had the value 0 and 1000 as construction year, which we agreed was an error. Given there were few observations with this error, we chose to drop them as it otherwise could have impacted the analysis.

There were some challenges to the data obtained from DMI. We observed that roughly 190 observations returned no information on all of the weather data collected from DMI. The missing observations where primarily for listings located in Copenhagen and Aarhus. Due to large markups on property in the larger cities it would presumably weaken our models accuracy at predicting listing in larger municipal areas if we dropped these observations. This issue was solved by taking the average values of the weather data variables and assigning to the missing data. This process salvaged roughly 190 observation which would otherwise have been dropped later on.

2.3.2 Time Shares

Going through the observations we came across some clear outliers with prices as low as 10.000 dkk and square meter prices as low as 1500 dkk. This required some investigation.
When scouting the listings on boliga.dk we found that some listings were not ownership of a property, but a part ownership of a timeshare. To improve our model we dropped all adverts with a price below 200.000 DKK, and a square-meter price below 6.000 DKK

2.3.3 Dummy variables and dropped observations

Next step was to create a dummy variable for parameters involving $city, selfsale, Municipality, andenergy_{c}lass.$ The reason behind this is that the regressive model can not take strings as an input. The dummies acts as a Boolean value; 1 if true, 0 if false. After creating dummies the former mentioned variables were dropped.

The data set contained some variables that were essential in the data generating process but irrelevant for the modelling.
Variables such as *Unnamed: 0 (a leftover index), id, links, Address, dawaID, sq_m_price* was dropped due to being irrelevant for the price valuation.*sq_m_price* was dropped as it reflects the valuation price and therefore the dependant variable.

After the previous cleanup we were still left with some missing or wrong values of which could not be corrected or saved. These values were then dropped, resulting in a complete data set of no missing values. Lastly we shuffled the data in order to remove any bias linked to the data generating process.

2.4 General descriptive statistics

2.4.1 House characteristics

Some general statistics and their distribution for the houses themselves.

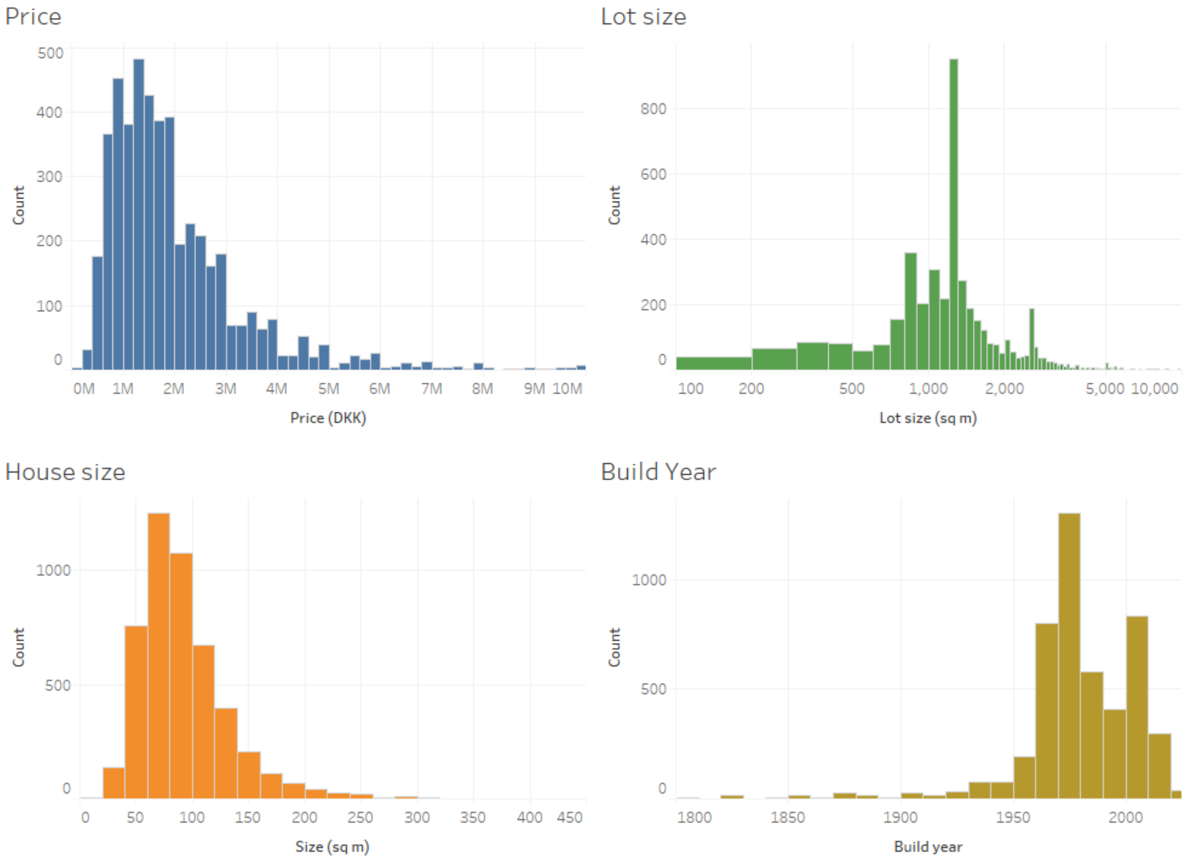
Table 1: Descriptive statistics

	size	price	lotsize	ejerudbetaling	buildyear	downPayment	n_rooms
mean	92.02082	2090566.60296	3299.84052	2062.76473	1978.30002	105522.59005	3.96814
std	45.65894	1720820.62775	24261.42610	1062.23393	28.83109	85817.16166	1.36064
min	11.00000	175000.00000	0.00000	0.00000	1777.00000	25000.00000	0.00000
25%	64.00000	1125000.00000	930.50000	1377.00000	1968.00000	60000.00000	3.00000
50%	83.00000	1685000.00000	1230.00000	1827.00000	1977.00000	85000.00000	4.00000
75%	109.00000	2495000.00000	1778.00000	2478.00000	1999.00000	125000.00000	5.00000
max	833.00000	25000000.00000	834191.00000	21303.00000	2020.00000	1250000.00000	20.00000

Figure 4 and table 1 show a right skewed distribution in prices and square meters, which indicates that most leisure homes are relatively small. Compared with the average size of family homes in Denmark at 140 square meters[2], it indicates that the trend for the leisure property market is smaller and cheaper. This is not unexpected, considering that leisure homes are secondary residences. Figure 4 also shows the distribution of lot sizes and build years. These are also right skewed. It is also noteworthy that the vast majority of houses were built in the 1960-70s, just before the oil shocks of 1973 and 1979, and in the

[2] Source based on BBR register: https://finansdanmark.dk/nyheder/2016/saa-meget-plads-har-danskerne/

Figure 4: Histograms for distribution of price, size, lotsize, and build year



Creation: Own
Data source: boliga.dk

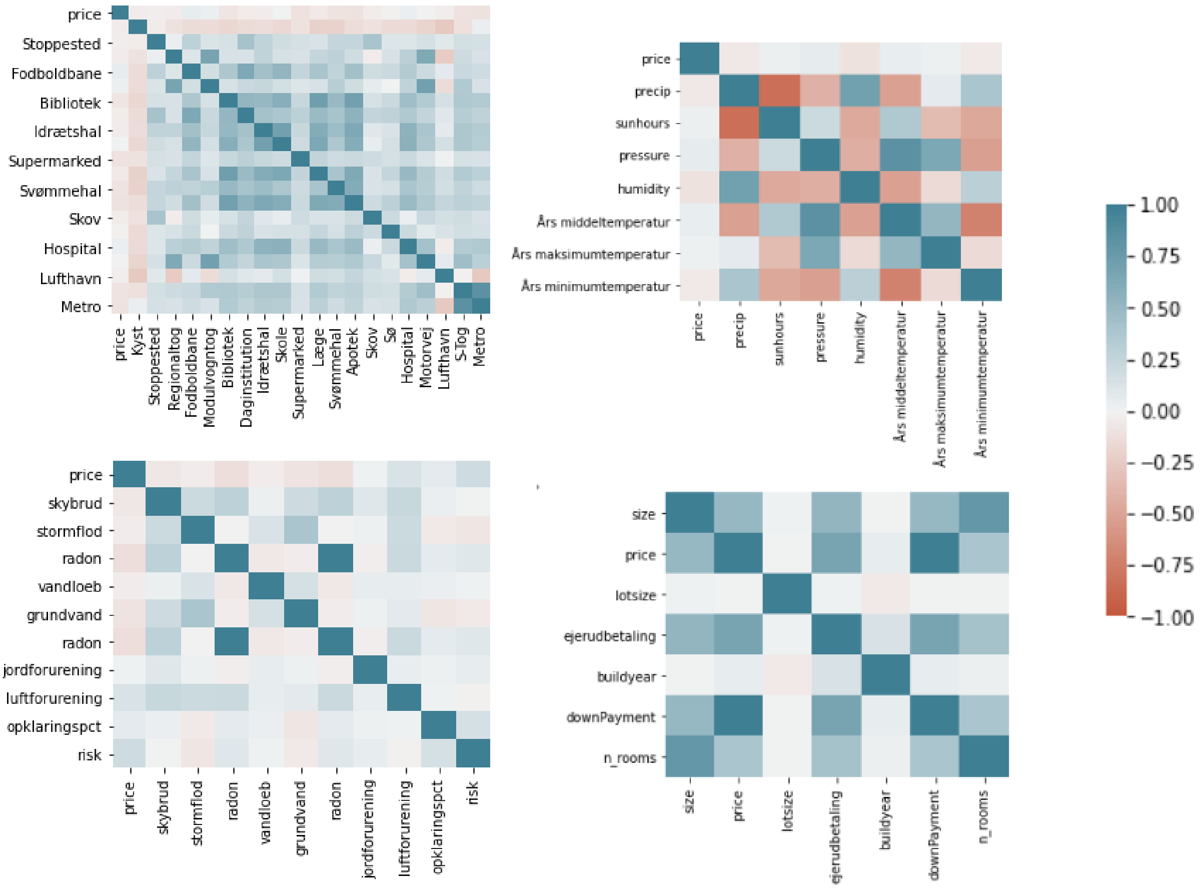
buildup to the financial crisis of 2008, when the housing market was running hot“.

The general statistics seems logical for the real estate market and general income distributions, as income and private fortune are right skewed as well across the population.

2.4.2 Correlation heat maps

In figure 6 we plot heat maps showing the correlation between the different variables and our target, price.

Figure 5: Correlation heatmaps for variables and target variable (price)



Source: boliga.dk, hvorlangterder.dk, dmi.dk/vejrarkiv

Most of our variables show a very weak linear correlation with the target variable (±0.1). However, the variables which present information available on the advert itself (price, size, down payment, number of rooms, owner expense) present a stronger correlation.

3 Machine Learning

3.1 Why use Machine Learning

Machine learning is in its essence the practice of letting computers or models recognize patterns and learn without specific programming, in order to produce reliable results and predictions. Machine learning models have been around since the 1950s, but it is only in recent years that it has been possible to combine machine learning with Big Data, at an increasing pace (SAS Institute, n.d.).

There exists several approaches for creating machine learning models. In this paper the supervised learning approach is used. The objective of the supervised learning approach is to train a model, with labeled data, to make reliable predictions on new data. Which in this paper is leisure home prices given several relevant features (Rashka & Mirjalli, 2017 p. 3).

3.2 Potential issues when applying Machine learning

Two common issues that occur when creating a machine learning model are overfitting and underfitting. Overfitting implies that the model captures the patterns in the training data too well, meaning it fails to generalise to new data. Underfitting implies the model is not flexible enough to capture the pattern in the training data. In both cases, the trained model fails to deliver reliable and accurate predictions. (Rashka & Mirjalli 2017, p. 73).

One simple way to cope with underfitting is to introduce more relevant features to the model, in order for the model to predict patterns and results on a more complex basis. Another is the introduction of polynomial terms, as these weaken the violation of linearity assumptions by generating polynomial features, resulting in a more precisely fitted model. Polynomial expansion should be used with care as it can result in overfitting (Rashka & Mirjalli, 2017 p. 334), and can quickly make the dataset too large to be processed.

Overfitting can be solved by finding a good bias-variance trade-off, by implementing regularization. Additional information is introduced in order to reduce the significance of extreme parameter values. Regularization can be implemented in two ways: Ridge and LASSO (Rashka & Mirjalli, 2017 p. 74).

3.2.1 Ridge

The Ridge regression adds the squared sum of the weights to the least-squared cost function. This means the variance and mean squared estimator will be smaller than previously, as the weights of the model is decreased, but never zero (Rashka & Mirjalli, 2017 p. 332). The Ridge regression is therefore useful when the parameters have more or less the same significance.

$$J(w)_{Ridge} = \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^{m}(w_j^2) \tag{1}$$

3.2.2 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) regularizes the weight so they can become zero (Rashka & Mirjalli, 2017 p. 332). The LASSO is applicable when there are few parameters that are significant.

$$J(w)_{LASSO} = \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^{m} |w_j| \tag{2}$$

One limitation of the LASSO is that it has a tendency to chose a variable from a highly correlated group and discard the rest of the variables in that group (Corporate Finance Institute, n.d.).

3.3 K-fold cross-validation

The objective of k-fold cross-validation is to minimize the mean squared error for the Ridge and LASSO regression, by optimizing the hyperparameter λ. The concept of k-fold cross-validation is that it is a 'resampling technique without replacement'. This implies that the training data is split into k folds without replacement. K-1 folds is then used for model training, while the remaining 1 fold is used for performance evaluation. This is repeated k times in order to obtain k models and performance estimates. The advantage of the k-fold cross-validation technique is that each observation is used for training and validation once, resulting in a lower variance estimate of the model performance (Rashka & Mirjalli, 2017 p. 191).

3.4 Model performance

Assessing the models performance can be done with a simple measure called Mean Squared Error (MSE), which is the average value of the Sum of Squared Errors (SSE). MSE is especially useful as it can compare different regressions model, since the SSE is normalized by the sample size (Rashka & Mirjalli, 2017 p. 330).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \tag{3}$$

3.5 Used model

The model used in this research paper is based on the pipeline procedure from Python. The dataset has therefore been split up into *development data* (2/3) and *test data* (1/3). Furthermore, the development data has been split into two equal sized subgroups, *training data* and *validation data.*
Based on the training data a model was 'trained' to predict the price, which was tested on the validation data. This exercise was done with both a LASSO and a Ridge regularization regression, in order to determine the model with the lowest Mean Squared Error. Based on the LASSO model a learning curve was made.

Secondly, a validation curve was made, which was based on the LASSO model also. This exercise is build on the development data. It obtains the MSE for different hyperparameter values and returns the one hyperparameter with the respective lowest MSE. Furthermore, a K-fold Cross-validation analysis was also conducted in order to further optimize the hyperparameter. This was done based on the LASSO model with 10 folds, as this is the usual standard.

4 Analysing results

4.1 Model Selection

In order to choose a model that performs the best given our data, an optimal hyperparameters is necessary. *Table 2* shows the optimal hyperparameters given one degree polynomial features for both LASSO and Ridge model.

Table 2: Optimal hyperparameters for one degree polynomial feature

	LASSO	Ridge
λ	1826.16	0.0093

Table 3: Mean squared error for Ridge & LASSO

	LASSO	Ridge
MSE	643,960,583.84	665,742,564.36
MSE CV	633,868,024.43	665,911,261.44
RMSE CV	25,176.74	25,805.26

The optimal lambda is 1826.16 and 0.0093 respectively for LASSO and Ridge. This indicate that the optimal degree in which the LASSO and Ridge model needs to penalize coefficients and weights is 1826.16 and 0.0093 respectively. The given optimal lambda makes it possible to compare the performance between the LASSO and Ridge. *Table 3* shows that the mean squared error (MSE) for LASSO and Ridge is 643,960,583.84 and 665,742,564.36 respectively. Taking the square root of these two numbers (Root Mean Squared Error, RMSE) returns 25,176.74 and 25,805.26 which indicates that LASSO performs better compared to Ridge for the given data set and the model. This implies that the model prediction is on average 25,176.76 DKK off from the true price.

4.2 Learning Curve

In order to assert for whether the model suffers from over- or underfitting problems and whether more data needs to be collected, a learning curve is examined. The learning curve of the chosen LASSO model is illustrated below in figure 6.

Figure 6: Mean performance LASSO

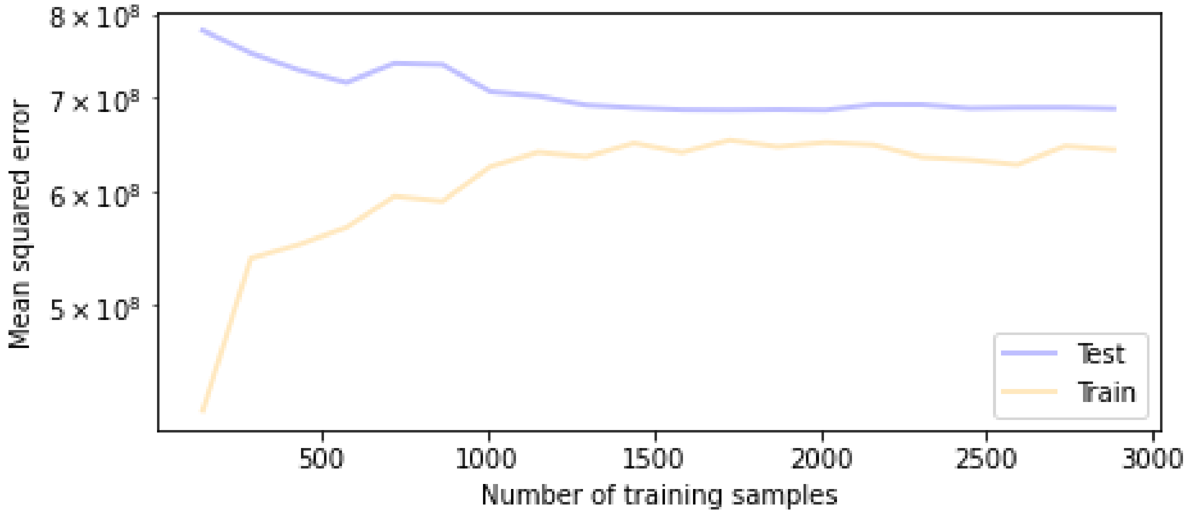


Figure 7 illustrates the average performance of both test and training set along with the number of training samples. At a low amount of training samples the distance between the test and training set is large. Although, the larger the sample size becomes the smaller this gap becomes. This indicates that the model performs well on both training and test dataset if there are more than 1000 samples during training. Moreover, the gap between the test and training set widens when the sample size is below 500 which indicates an increasing degree of overfitting.

4.3 Validation Curve

In order to improve the performance of the selected model, a validation curve was also constructed to address the issue of over- or underfitting. Illustrated below is a validation curve of the LASSO model based on the train and validation data set (see figure 8).

Figure 7: Validation curve LASSO

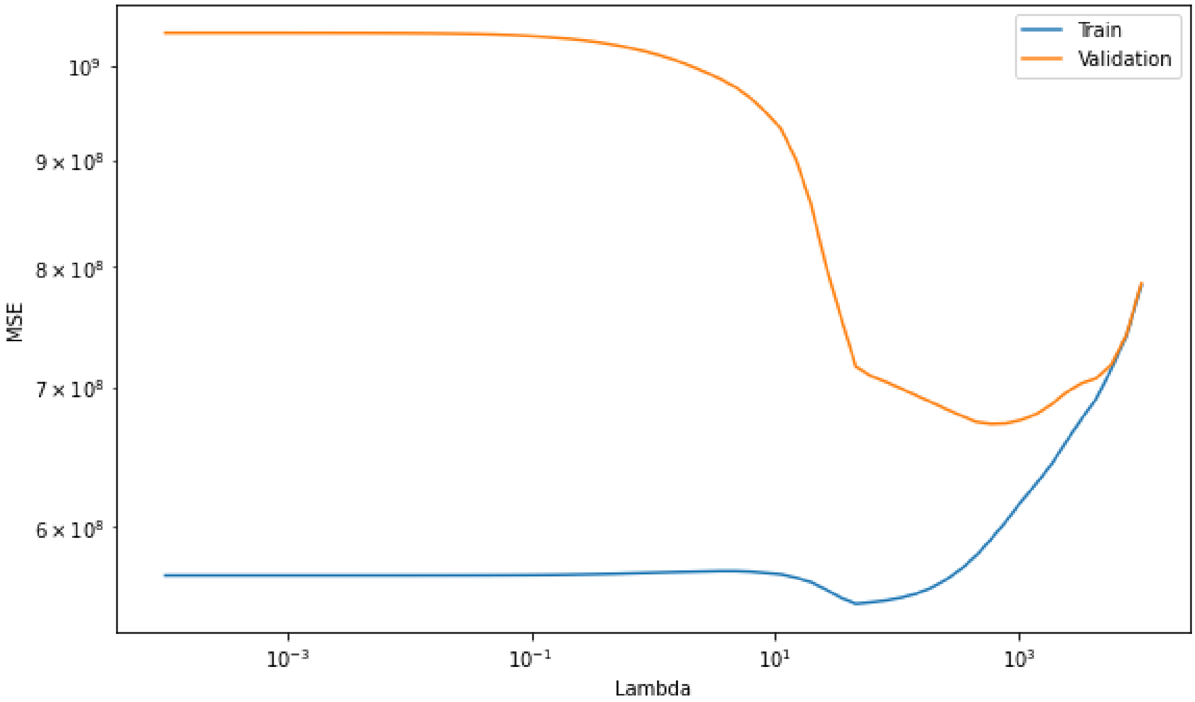


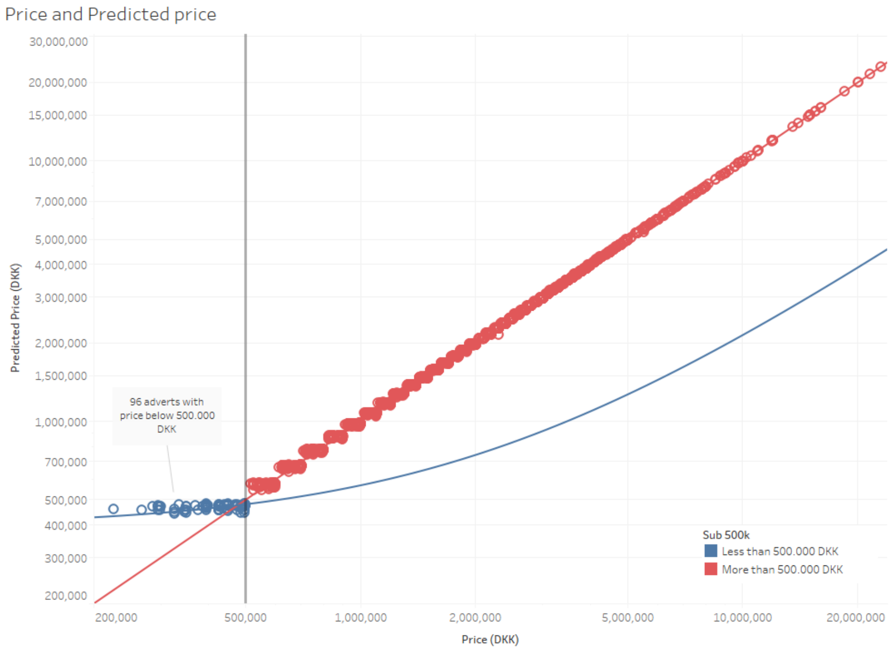
Figure 7 illustrates the average performance of training and validation set along with the corresponding lambda values. Similar to the learning curve the distance between the two set is very large when lambda is small. This suggest that if the chosen lambda was small, implying decreasing the strength of regularization, the variance between the two set is large and the model is overfitted. On the other hand, choosing a large lambda, meaning increasing the regularization, the training and validation performance becomes horrible. This is due to the coefficient being penalized so much that they become close to zero. This suggests that at higher values of lambdas, the model would be too biased and tend to be underfitted. We can observe that the model performs better on the training data compared to the validation data. This suggests that the model tends to be overfitted. One way to adjust for this is to increase the number of samples on the training data. Although, *figure 7* showed that the variance increases a bit around 2000-2500 samples. This indicates that increasing the number of samples on the training data would not have any effect on decreasing the variance. Another approach would therefore be to adjust for hyperparameters in order to minimize the predictive errors.

Overall, the results suggests that the optimal hyperparameter has been found which yields the balance between bias and variance. Moreover, the validation curve is to some degree, fairly smooth. This indicates that the performance of the validation curve is not random and adds to the credibility of the optimal hyperparameters. Hence, the model has been optimize given the available resources.

4.4 Model Performance Results

The overall performance of our model is satisfying with a mean deviation of $\sim$ 25.000 DKK (about 1.2% of the mean price) from the listed prices on boliga.dk. The following figure 8 shows predicted vs actual valuation, the two trendlines shows trends for property listings above and below 500.000 DKK.

Figure 8: Predicted price Vs True price

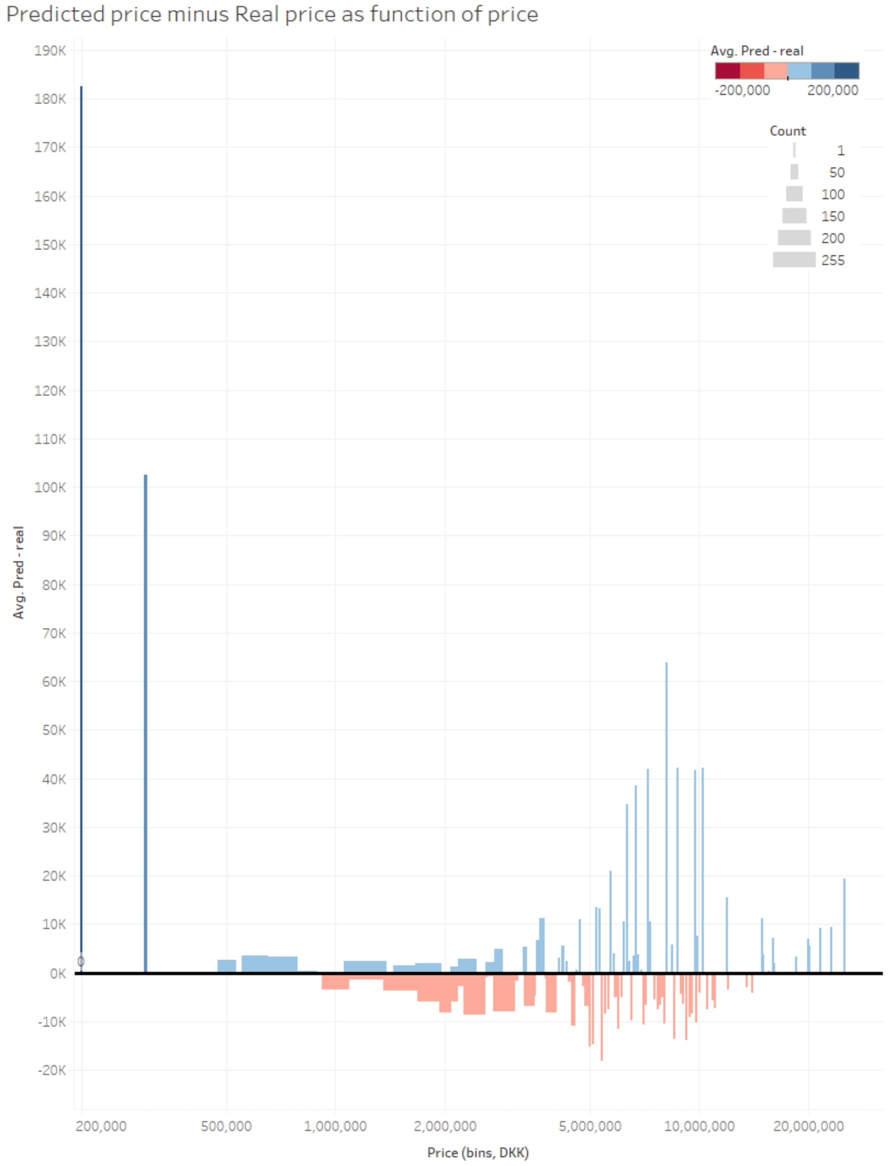


Creation: Own
Data source: boliga.dk, own model

Figure 8 highlights the accuracy of our model: The overall R-squared for a linear fit (not shown) is of 0.9993, where a value of 1 would be a perfect prediction for all examples. Looking at this figure alone, it would seem that the model is particularly bad at predicting houses with a low price (below 500.000 DKK). The R-square for a linear fit of these 96 observations (in blue) is of only 0.389. The remaining predictions follow the real values almost perfectly: The R-square for this group (shown in red) is even higher than the overall, with a value of 0.9999. However, we still observe a clustering around each 100.000 DKK mark. At first glance, this would suggest the model is not actually better at predicting prices for homes with a price above 500.000 DKK either.

To examine this more in detail, we plot the average (per bin) prediction error as a function of price in figure 9. Here we can clearly see how well our model works: For a price below 500.000 DKK, the prediction errors are large in absolute value, but there are few observations in this range. For a price above 2 million DKK, the predictions are more accurate, but less precise: absolute value error is small, but varies highly. The model has the best performance in the price range of 1 million to 2.5 million, where it is both highly accurate and highly precise: Low absolute prediction error with low variance. We can explain this by looking back at table 1 and figure 4: This range corresponds to the interquantile range, where the bulk of the data is found. As we could have expected, our model's strength follows the skewed

Figure 9: Predicted price minus Real price, as function of price
(NB: y-axis is asymmetric around y=0 to better show the estimation errors)



Creation: Own
Data: boliga.dk, own model

distribution of prices, which itself reflects the general socioeconomic distribution in most countries. For the range with the most observations (indicated by the bar width), the model performance is best, and the prediction error is much higher for the rest of the observations, which is caused by the way our data set is biased.

One further consideration is the fact that while the model is inaccurate in absolute values for more expensive homes, this shortcoming disappears if we consider prediction error relative to price. This is shown in figure 10 in the appendix. It is clear here that the relative error is of less than 0.2% on average for a price above 2.5 million. **Overall, our model overvalues houses by 1.8% of their actual price on average, or about 25.000 DKK in absolute terms.**. We are very satisfied with this performance which we consider to be extremely strong.

5 Discussion

5.1 Data Limitation

A different approach could be to use the same method, but on sold leisure houses to predict the selling price. The predicted price setting and predicted selling price could be compared to see the potential revenue. It can be argued that the real estate agent is mostly interested in the actual selling price instead of the valuation price. Having information about the actual selling price in an apartment complex or street can aid the price setting of other leisure houses with similar characteristics.

Another limitation to the data is non parametric variables such as the view or wildlife in the area. These data was not present on the websites that was scraped and may have an influence on the price setting on leisure houses. Some may choose to rent a leisure house in a forest because they value nature and wildlife more than being close to a beach. This is not captured in our model and may be a significant unseen variable for leisure homes specifically.

The variables describing health risks are not necessarily correct for the individual property, as they are calculated on area statistics. For an elaborate description of the individual properties it would be necessary to access individual status reports on every single property, which are not publicly accessible.

5.2 Model Limitation

Implementing a time series model might have showed a trend on how the prices move as well as what may influence the prices e.g. weather or shocks like a pandemic. Furthermore, will a times series model show the demand for leisure houses over time and their incentives to purchasing one.

Due to the the specific approach of predicting valuation based on current listings and not development over time, we are limited strictly to geographical features in our model. It would have been interesting to model how developments in the average income, private equity or general economic cycles impact the valuation. However if one is to include variables such as income in the different area codes, it would not reflect well on a model covering leisure properties. The average income in an area is based on the individuals home address. This means that sociodemographic data would not be representative for the areas of interest for leisure properties, as most buyers do not live right next to their leisure home. In terms of data on general economic development, it would be redundant as the modelling consist of a single period. However, for time series analysis it would be relevant to analyse how variables such as average: $\frac{property\ expenses}{income}$ or GDP growth affect the valuation.

An alternative to the use of the LASSO regularization method is the Elastic Net. The Elastic Net combines the Ridge and LASSO regularized linear regressions hereby including the two penalties, L1 & L2. Using the Elastic Net some of the limitations (number of selected variables) of the LASSO are eliminated, since the Elastic Net includes "n" number of variables (Rashka & Mirjalli, 2017 p. 333).

$$J(w)_{ElasticNet} = \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \lambda_1 \underbrace{\sum_{j=1}^{m}(w_j^2)}_{L1} + \lambda_2 \underbrace{\sum_{j=1}^{m}|w_j|}_{L2} \tag{4}$$

5.3 Model precision

It would have been relevant to consider an analysis of insignificant variables for a cleaner data set. As the LASSO performed best, it could indicate the presence of insignificant variables in our data set, as we found with the heat map of correlation with the dependant variable.

Furthermore, we observe that a large amount of the errors between the modelled price and actual listed price comes from 96 observations valued at less than 500.000 DKK and for valuations above 4.000.000. These errors reflect that the model might be biased by the majority of observations being listed between 500.000-3.000.000. The model might have benefited from a more uniform distribution which is not reflective for the market, or by dropping observations in the really cheap and expensive price ranges.

Originally the intention was to include leisure status building sites in order to train the model for the raw value of land. However, we faced the obstacle of too few listings in order to train an accurate model. A plot of the empty properties showed some areas with none or few listings, which would not be enough for training the model. This is a huge drawback for our model, as the price of land should be a relatively common valuation strongly correlated with location and geographical features.

Lastly, we believe that one reason for our model being strong is the relatively narrow dataset used, focusing exclusively on leisure homes. Had we tried to predict house prices in general,it is probable that our model performance would have been inferior, considering the much larger variability in features. This would also have required more variables regarding socioeconomic factors to make an accurate prediction concerning the price of a permanent residence, rather than a secondary residence.

5.4 Ethical Challenges

The advancement in technology has lead to the accessibility of big data. This has led to privacy frameworks for both the public and private-sectors regulating how this data can be used in practice. The protection of privacy is important to ensure the security of individuals, as well as the the fair and ethical use of their information. These privacy principles are challenged by big data in the sense that the individual has limited control over their personal data and the processing thereof. Furthermore, meaningful informed consent is another issue when working with big data due to the involvement of continuous collection of data over time when analysing big data. The data is only accurate as long as it's up to date (Lacroix, 2019).

Another ethical challenge is what sort of information may be publicly accessible. Information through e.g. Facebook or Twitter's API may be valuable to some industries which purpose includes market research, advertisement or background demographics. Connections and social network may be reconstructed through these information and for the first time makes it possible to observe large-scale social interactions in real time (Neuhaus & Webmoor, 2012).

Most electronics devices nowadays have an integrated location-based software which gives rise to another ethical consideration. Access to this sort of data gives issues to the perception of public and private space of the individual. The idea of being monitored without knowing is a major breaking of privacy and could potentially harm the individual (Neuhaus & Webmoor, 2012). To protect the individuals The European Union (EU) has implemented a legislation called General Data Protection Regulation (GDPR) in May 2018. It strengthens and harmonizes the protection of personal data for citizens in the EU. The GDPR takes into consideration whether the location or the personal data is being processed regardless of where the data controller is located in the world. This legislation demonstrates accountability when collecting and using personal information. In terms of big data analytics the GDPR mainly focuses on

how the data is used and not necessarily how it was collected (Lacroix, 2019).
The modern age of technology has increased the possibility of generating useful data on consumption and behavior, which improves efficiency in private and public sectors. However, the usage and degree of collecting this data becomes a trade off between surveillance and benefit.

6 Conclusion

This research paper has made an attempt to use machine learning to train a model to predict prices for leisure houses in Denmark. The model was trained on geographical and weather condition data. These data was scraped from Boliga.dk to obtain lists of houses, hvorlangerder.dk returned the distances to different every day commodities such as supermarkets and pharmacies, and DMI.dk showed the average weather patterns for different municipalities. A LASSO and Ridge model were then trained on the data and the performance of each model was compared. Through a k-fold cross-validation the optimal hyperparameter λ was obtained for both models. This resulted in the LASSO model performing better compared to Ridge when predicting house prices with a MSE of $25,176.74^2$ and optimal hyperparameter of 1826.16.

In order to address the over- and underfitting problems a learning and validation curve was examined. These curves showed that increasing the sample size would not lead to a decrease in variance between the training and test dataset. As for the validation curve does the LASSO model tend to overfit the dataset as it performs better on the training set compared to the validation set. However, the results indicates that the optimal hyperparameter has been found to yield a good balance between bias and variance which adds to credibility.

Lastly, the results as well as the limitations of the model and data were discussed, as well as alternative approaches such as implementation of time series model that could be considered for future research. In conclusion, given the chosen features, we were able to train a prediction model whose performance we were very satisfied with. On average, the prediction error relative to price was of 1.2%, or an absolute error of roughly 25,000 DKK compared to the average price in our dataset, while the model overvalues by 1.8% of price on average.

7 Literature

Chen, Philip & Zhang, Chun-Yang (2014): Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Volume 275, Pages 314-347, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2014.01.015.

Corporate Finance Institute (n.d.): *Elastic Net*, [online]. Corporate Finance Institute. [23-08-2021]. Available: `https://corporatefinanceinstitute.com/resources/knowledge/other/elastic-net/`

Fan, G-Z. Ong, SE & Koh, HC. (2006): Determinants of House Price: A Decision Tree Approach. Urban Studies. 43(12):2301-2315. doi:10.1080/00420980600990928

Lacroix, P. (2019): Big Data Privacy and Ethical Challenges. In: Househ M., Kushniruk A., Borycki E. (eds) Big Data, Big Challenges: A Healthcare Perspective. Lecture Notes in Bioengineering. Springer, Cham. $https: //doi.org/10.1007/978 - 3 - 030 - 06109 - 8_9$

Neuhaus, Fabian & Webmoor, Timothy (2012): AGILE ETHICS FOR MASSIFIED RESEARCH AND VISUALIZATION, Information, Communication & Society, 15:1, 43-65, DOI: 10.1080/1369118X.2011.616519

Park, Byeonghwa & Bae, Jae Kwon (2015): Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, Pages 2928-2934, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2014.11.040.

Rashka, Sebastian; Mirjalli, Vahid (2017): *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow.* Second Edition. Packt Publishing, Mumbai

Salganik, Matthew J. (2017): *Bit by Bit: Social Research in the Digital Age.* Princeton, NJ: Princeton University Press. Open review edition.

SAS Institute (n.d.): *Machine Learning: What it is and why it matters*, [online]. SAS Institute. [19-08-2021]. Available: `https://www.sas.com/en_us/insights/analytics/machine-learning.html`

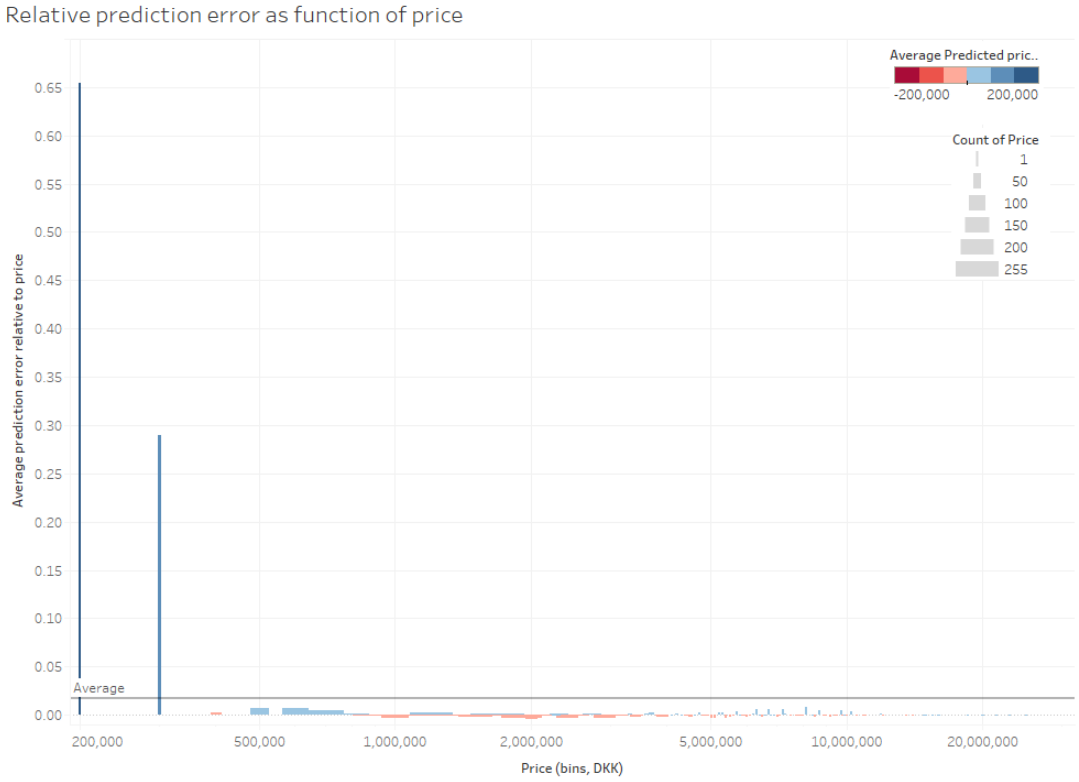
Selim, Hasan (2009): Determinants of house prices in Turkey: Hedonic regression versus artificial neural network, Expert Systems with Applications, Volume 36, Issue 2, Part 2, Pages 2843-2852, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2008.01.044.

Winky, K.O. Ho, Bo-Sin Tang & Wong, Siu Wai (2021) Predicting property prices with machine learning algorithms, Journal of Property Research, 38:1, 48-70, DOI: 10.1080/09599916.2020.1832558

Yaqoob, Ibrar. Hashem, Ibrahim Abaker Targio. Gani, Abdullah. Mokhtar, Salimah. Ahmed, Ejaz. Anuar, Nor Badru & Vasilakos, Athanasios V. (2016): Big data: From beginning to future, International Journal of Information Management, Volume 36, Issue 6, Part B, Pages 1231-1247, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2016.07.009.

8 Appendix

Figure 10: Prediction error relative to price, as function of price



Creation: Own
Data: boliga.dk, own model

Table 4: Data set variables - not including dummies

Variable	Source	Included in model	Class	Description
id	boliga.dk	False	int	Boliga.dk database ID
link	boliga.dk	False	str	Boliga.dk url
adress	boliga.dk	False	str	Advert address
zip	boliga.dk	False	int	Danish zip code
city	boliga.dk	False	str	Advert city
kommune	boliga.dk	False	str	Advert municipality
dawaID	boliga.dk	False	str	Danish GPS ID
latitude	boliga.dk	False	float	Latitude
longitude	boliga.dk	False	float	Longitude
sq m price	boliga.dk	False	float	Square metre price
price	boliga.dk	TARGET	float	Listed price
size	boliga.dk	True	float	House size
lotsize	boliga.dk	True	float	Lot size
ejerudbetaling	boliga.dk	True	float	Owner expense
buildyear	boliga.dk	True	int	Build year
energy class	boliga.dk	True	categorical	Energy class
downPayment	boliga.dk	True	float	Down Payment
n rooms	boliga.dk	True	int	Number of rooms
selfsale	boliga.dk	True	boolean	Sold by owners without an agent
skybrud	boliga.dk	True	float	Risk associated with heavy rain
stormflod	boliga.dk	True	float	Risk associated with sudden wind-caused sea level rises
vandloeb	boliga.dk	True	float	Risk associated with rain-caused rise in water bodies levels
grundvand	boliga.dk	True	float	Risk associated with water contained in the ground
radon	boliga.dk	True	float	Risk associated with the radioactive gas radon
jordforurening	boliga.dk	True	float	Ground pollution
luftforurening	boliga.dk	True	float	Air pollution
opklaringspct	boliga.dk	True	float	% of burglaries cases solved by police
risk	boliga.dk	True	float	Risk of burglary
precip	dmi.dk	True	float	Average yearly quantity of rainfall (mm), 2010-2021
sunhours	dmi.dk	True	float	Average yearly hours of sun, 2010-2021
pressure	dmi.dk	True	float	Average yearly air pressure (hPa), 2010-2021
humidity	dmi.dk	True	float	Average yearly air humidity (%), 2010-2021
Års middeltemperatur	dmi.dk	True	float	Average yearly temperature (°C), 2010-2021
Års maksimumemperatur	dmi.dk	True	float	Average maximum yearly temperature (°C), 2010-2021
Års minimumltemperatur	dmi.dk	True	float	Average minimum yearly temperature (°C), 2010-2021
Hvorlangterder variables	hvorlangterder.dk	True	float	Distance to nearest point (see translated list)
kyst	hvorlangterder.dk	true	float	distance from coast
stoppested	hvorlangterder.dk	true	float	distance from bus
regionaltog	hvorlangterder.dk	true	float	distance from regional train
fodboldbane	hvorlangterder.dk	true	float	distance from soccer field
modulvogntog	hvorlangterder.dk	true	float	distance from train
bibliotek	hvorlangterder.dk	true	float	distance from library
daginstitution	hvorlangterder.dk	true	float	distance from day care
idrætshal	hvorlangterder.dk	true	float	distance from gym facilities
skole	hvorlangterder.dk	true	float	distance from school
supermarked	hvorlangterder.dk	true	float	distance from groceries
læge	hvorlangterder.dk	true	float	distance from doctors office
svømmehal	hvorlangterder.dk	true	float	distance from swimming facilities
apotek	hvorlangterder.dk	true	float	distance from pharmacy
skov	hvorlangterder.dk	true	float	distance from Forest
sø	hvorlangterder.dk	true	float	distance from lake
hospital	hvorlangterder.dk	true	float	distance from hospital
motorvej	hvorlangterder.dk	true	float	distance from highway
lufthavn	hvorlangterder.dk	true	float	distance from airport
s-tog	hvorlangterder.dk	true	float	distance from s-train
metro	hvorlangterder.dk	true	float	distance from metro