UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS

Location Dynamics: Job Markets and Transit Access in Danish Property Value Models

Can we improve property value models by including features for job market and transit access?

Date: August 28. 2024

Group: 20 (Jesper Bork Petersen, Tobias Thiim Mølgaard & Nicoline Lund Dahl)

Number of characters: 35,431

Contributions from each group member is listed in the table of contents.

Table of contents

1. Introduction[1]

Understanding the dynamics of real estate prices is of use not only for individual buyers but also for a wide range of stakeholders, including researchers, urban planners, and policymakers. Effective modelling of these dynamics provides valuable tools for analysing market patterns, which in turn enables urban planners to promote sustainable development and empowers policymakers to shape resilient real estate markets (Kang et al., 2021).

Our small study aims to contribute to this body of knowledge by integrating key area-specific characteristics concerning professional opportunities—such as job opportunities and public transportation accessibility—into the analysis of housing prices within Denmark's unique urban landscape.

Our Research Question is: *To what extent can area-specific factors* concerning professional opportunities *improve predictive models of property values in Denmark*? We seek to examine the relation between job market dynamics, public transportation and real estate prices by employing a Machine Learning approach to the problem.

We answer this research question by collecting housing data from Denmark's largest real estate search engine. To enrich this data, and investigate whether area-specific factors that enhance professional opportunities improve our ability to predict prices, we further scrape job listings from a state-run job portal as well as information about the nearest train station and number of departures from DSB and Banedanmark.

We then train a series of Machine Learning models on the data and compare the results. We look at Lasso, Ridge, Elastic Net and Random Forest. We arrive at the conclusion that the extended model with features for *area-specific factors* makes more accurate out of sample predictions compared to the more simple model which only takes classical real estate features such as property size, number of rooms, municipality etc. This is true across all four types of regularization. Overall, the extended Random Forest performs best.

[1] We used ChatGPT as a tool in this assignment, primarily for code refinement and text editing. Specifically, it assisted in amending code and in shortening text during the revision process to reduce word count. The assignment was co-written and co-edited collaboratively, with random attribution of sections to meet formal requirements.

2. Literature Review

The housing market has for a long time been of major interest in the field of economics due to the market's strong impact on the economy, and as a potential driver for inequality in capital (*Andersen, 2021*).

In recent years a number of articles have looked further into the relation between housing prices and spatial characteristics. *Agnew & Lyons (2018)* denote access to employment as a 'key determinant in house pricing'. Their paper examines the link between employment and Irish housing prices using spatial datasets on housing prices and employment. *Kim and Jin (2019)* investigated how improvement of job accessibility and mixed land use affect housing prices. Using an instrument variable design *Kim and Jin (2019)* find that increased job accessibility increases house prices, whereas mixed land use decreases house prices.

Gibbons and Machin (2005) have investigated the relation between access to railway stations and housing prices. Using a quasi-experimental approach, they find that improved access to railway stations have a large effect on housing prices compared to the valuations of other local amenities.

Rojas (2024) argues that public infrastructure investments are often motivated by a desire to improve accessibility. Increased accessibility in turn, as per *Gibbons and Machin (2005)* manifests in higher house prices, in so far as markets are efficient. *Rojas (2024),* further shows that the empirical evidence is decidedly mixed – and far from unequivocal on this issue. Multiple studies have – perhaps counterintuitively - found that proximity to new metro or light rail stations can in fact have a negative sign on real estate prices. In *Hess and Almeida (2007)* this occurs when neighborhoods are classified as low-income. In a study by *Zhou et al. (2021)* the effects of increased transportation accessibility appear to interact strongly with the level of proximity to employment centers.

The term "employment center" (EC) refers to "a site of significant geographic concentration of economic activity" *(Giuliano & Redfearn, 2005)*. Several definitions exist for identifying ECs, all aimed at describing locations with a significantly higher concentration of jobs compared to surrounding areas, making them stand out as hubs of employment activity. ECs can be defined in both relative and absolute terms. *Giuliano and Small (1991)* propose an influential measure, which, when converted to square kilometers, defines an EC as a cluster of adjacent zones where each has a job density exceeding 2,471 jobs per square kilometer and collectively contains more than 10,000 jobs.

Often, previous studies have found that there is a high degree of spatial autocorrelation when it comes to evaluations. Recently, research into determinants of real estate prices is making increased use of machine learning. For instance *Chao et al. (2020)*e, who define three urban public transportation indexes and exploit these to train a Random Forest ML model which predicts real estate prices in Xi'an, China.

3. Research Design

In the following we will shortly describe our research design. Based on the literature, we expect that area-specific characteristics enhancing professional opportunities are important factors when estimating real estate prices. We therefore want to test whether a machine learning model that takes these types of area-specific characteristics into account can make more accurate predictions, than a machine learning model that only takes classical features such as number of rooms, postal code, type of property etc. into account.

We conceptualize area-specific characteristics that enhance professional opportunities as job opportunities close by or mobility that enables the individual to easily travel to and from the workplace.

To operationalize job opportunities we use a grid, to divide Denmark into smaller cells of 10 by 10 km. We then calculate the job density within each cell. All properties for sale are situated within a specific cell, and we can then add the job density within the given cell as a feature of each property. A disadvantage with the grid method is that even though your property is situated in a grid cell with low job density, you could potentially live just a few meters from a grid cell with high job density.

To mitigate the risk of overlooking high job density in a grid cell near the property, we instead create a distance based measure that calculates the distance to the nearest cell with very high job density. Practically we identify the 25 grids with the highest job density and define those as *Employment Centers* (*EC*). We then calculate the air route distance to the nearest employment center. We hereby end up with two different measures for job opportunities close by (i) job density in grid (ii) distance to nearest EC.

An alternative approach could be to count the number of jobs within a 5-, 10- and 20-kilometer radius of each property. This alternative method is however discarded due to computational complexity.

We operationalize mobility as the air route distance to nearest railway station. Using the air route distance has some drawbacks, especially in a country like Denmark that has a lot of islands, because the distance to the nearest station can be calculated over water even though the actual travel route may be much longer and include a ferry crossing. Air route distance is however chosen due to computational simplicity.

To further quantify the quality of the station, we add a second feature which is the number of departures from the nearest stations between 08.15 and 09.00 a Thursday morning mid-august. This time point is chosen because it indicates opportunities to get to work using the mobility facilities.

4. Data

This section explains how data on properties for sale, job postings and information on train stations and departures were collected. The aim of the section is to describe the process as well as to give an insight to the information each dataset contains.

4.1. Data Ethics

To address our research question, we did not use any personal or individual data, so no participant consent was necessary. However, we carefully considered ethical aspects before scraping to ensure compliance with laws and to avoid harm.

First, all data used in this study is publicly accessible, making scraping a method of automating data collection within ethical boundaries.

Second, we checked for available APIs before scraping and used them whenever possible instead of extracting HTML code.

Third, we identified ourselves through user-agent strings and provided contact information, allowing site owners to reach out with any questions or concerns.

Fourth, we controlled our data requests with a sleep timer to avoid overwhelming the servers and prevent any confusion with a DDoS attack.

Lastly, we only stored and used the data necessary to answer our research question.

4.2. Boliga

Boliga.dk is a Danish online portal which has existed since 2007 and contains information on all properties for sale in Denmark, e.g. price, location, type of property etc.

4.2.1. Data collection

Using the DevTools module in Google Chrome we identified an API on all properties for sale. We therefore used this to collect data from Boliga.dk In total 55,265 observations were collected. The data was collected on the 16th of august 2024.

4.2.2. Data Cleaning

Figure 4.1 gives an overview of the data cleaning process on real estate data. Each step is described in further detail below.

Figure 4.1: Cleaning process for property data

Data collection from Boliga.dk (55,265)

Data Cleaning

Description	Number of properties	Change
Data from Boliga.dk	55,265	
1. Exclusion of observations with missing coordinates	55,257	-8
2. Exclusion of cooperative housing, holiday homes and lots	43,429	-11,828
3. Exclusion of properties with missing price	43,428	-1
4. Exclusion of properties with a size at 10 sqm or less	43,386	-42
5. Exclusion of garages	43,384	-2
6. Exclusion of duplicate values	43,329	-55
7. Exclusion of observations with missing values in key statistics	42,709	-620

(1) First we remove 8 observations with missing coordinates.

(2) Cooperative housing, governed by legal and cooperative pricing mechanisms, differ significantly from owner-occupied listings. Similarly, holiday homes are influenced

by distinct factors. Given our focus on the impact of jobs and transport on property prices, these properties have been excluded from the dataset.

(3) We excluded one property with a price of zero.

(4) We also removed 40 properties sized 10 sqm or smaller since these were all commercial properties. However, it should be noted that this exclusion does not guarantee the removal of all commercial properties.

(5) 2 garages were excluded.

(6) 55 duplicate entries were removed from the data..

(7) Finally, we excluded 655 observations with missing values in one or more key statistics. We did this to obtain a balanced dataset ready for Machine Learning. We also excluded 4 observations with extreme values of expenses.

After these steps we have a dataset consisting of 42,709 observations.

Because some apartments had a lot size at zero and other apartments had a lot size corresponding to that of the entire property, we set the size to zero for all apartments. Further we decided to unite the property types 10, 11 and 12 as they all correspond to “Other type”.

4.3. Jobnet

4.3.1. Scraping

To obtain information about jobs we scrape jobnet.dk. Jobnet is a state-run search engine, which aggregates both public and private job databases. The website falls under the Danish Agency for Labour Market and Recruitment.

The scrape itself is simple[2]. We encounter no limits and are able to run the code without problems. The data was scraped on the 15th of august 2024.

[2] The code looks much what we practiced in Exercise 6

We end up with a dataset with 17,980 rows × 46 columns. The data contains characteristics of each job e.g. occupation area, work hours and a job description. Importantly, we get the location coordinates of the workplace.

4.3.2. Data Cleaning

In order to obtain accurate and consistent data we delete duplicates and drop entries that lack location values.

When inspecting the job data, we find that all jobs from the municipality of Copenhagen are located at the address of the *Koncernservice* department, even though the actual workplaces are schools, youth centers, nursing homes etc. spread out all over the municipality. Because institutions such as schools and youth centers are pretty evenly distributed all over the municipality of Copenhagen, we chose to randomly assign locations within the municipality instead of keeping all the job postings at the address of the *Koncernservice* department.

Finally, we remove jobs placed outside of Denmark, in order to only map distances relevant to the real estate listings in our analysis. After this data cleaning process, we have 17,643 job entries.

4.4 Station data

To obtain information on Danish stations we scrape information on stations and their coordinates from the website of the state-owned company Danske Statsbaner (DSB). This provides us with a list of 320 railway stations in Denmark. We cross-refer the list of stations from DSB with those available at the 'MitTog' website which is hosted by the state-owned company Banedanmark. By doing this we find 7 stations which are not actual railway stations but instead long distance bus stations primarily situated on danish islands. We remove these stations from our dataset, which leaves 313 railway stations. Metro stations without other connections are not included in the list of stations.

To obtain a measure regarding the number of departures from the nearest station we scraped station data from the MitTog website. Practically, we used the selenium package to navigate to the pages of the specific stations, and then scrape the departure boards. Because the website didn't have a time selector, we had to visit the website at the desired time point. The code however needed some time to run, because each of the dynamic departure boards had to be fully loaded, before we could scrape the data. This implied that we couldn't visit each of the station pages at exactly 8.15. We therefore visited the station webpages in the period from 8.08 until 08.22. We then calculated the number of departures in the following 45 minutes, from when we visited the specific webpage. This method however, implies a small risk of

bias, if trains from Danish stations in general leave more often in the timespan between 08.08 and 08.20 than in the timespan between 08.53 and 09.05. If this is the case, we could systematically assign a lower number of departures to stations starting with a letter coming late in the alphabet. We however expect these differences to be marginal, and we do not expect a strong systematic order of starting letters.

In general we hereby ended up with a dataset containing 313 stations, their specific coordinates, and the number of departures in 45 minutes a weekday morning in august.

4.5 Data wrangling

In order to combine the three datasets, we exploit the fact that all datasets contain geographic coordinates. We start off by creating the grid, dividing all of Denmark into separate square cells of 10x10 km. We choose 10x10km cells over larger ones[3] given the relatively compact size of Denmark.

We then convert both the job and the real estate datasets to GeoDataFrames (gdf), by using the GeoPandas (gpd) package in Python. This allows us to count how many jobs that fall into a specific cell. For each property in the real estate dataset, we then identify the cell in which the property is situated.

To further determine the distance to the nearest employment center we filter out the 25 cells with highest job density and then for each property, we use the KDTree package to calculate the air line distance to the nearest employment center in kilometers. This number is then defined as a feature of each real estate.

What we do with stations is very similar. We use the KDTree package and our station dataset to calculate air line distance from a given property to the nearest station, and ascribe number of kilometers as a feature of the listing. Further, we identify the nearest station and take the number of departures from the station dataset and add it to the real estate dataset as a feature of the listing.

After performing these data wrangling steps, we have a dataset containing information on each real estate and nearby job and infrastructure possibilities.

[3] Bitzer & Goren, 1998, in a development economics study use 0.5 decimal degrees latitude × longitude, approximately 55 km × 55 km.

5. Visualization and Descriptive Statistics

This section presents visual and descriptive statistics of the final dataset.

5.1. Key Statistics of numerical variables

Table 5.1 presents key statistics of the numerical variables in the final dataset.

Table 5.1 Key statistics of numerical variables

		count	mean	std	min	25%	50%	75%	max
Outcome	Price (mill.)	42,709	3.13	3.71	85	1.295	2.195	3.7	120
Property variables	Rooms	42,709	5	2	1	4	5	6	71
	Size (sqm)	42,709	151	66	17	109	141	180	1,268
	Lot size (sqm)	42,709	5,596	41,209	0	400	800	1,164	4,119,367
	Build year	42,709	1948	48	1575	1920	1959	1976	2026
	Monthly expenses	42,709	2.783	2,292	90	1,578	2,207	3,269	69,235
	Basement size (sqm)	42,709	27	193	0	0	0	15	23,708
Job variables	Job density	42,709	134	276	0	4	22	103	1,432
	Distance to Job Center	42,709	25	24	0	5	18	35	171
Station variables	Distance to nearest station	42,709	8	16	0	1	3	10	171
	Departures per hour	42,709	5	7	0	2	3	5	63

Table 5.1 shows that both the outcome variable (*price*) and all input variables, except *Build year*, are right-skewed. The median property in our dataset costs 2,195,000 DKK, is 142 sqm, has 5 rooms, is situated on an 801 sqm lot, was built in 1959, and has a monthly expense of 2,204 DKK. The maximum price is 120,000,000 DKK, and the smallest property is a 17 sqm apartment.

Our dataset includes 7,172 properties with a lot size of 0 sqm—all apartments, since we correct the lot size to zero for all apartments, as outlined in section 4.2.2. The largest lot, 4,119,367 sqm, is a country estate. Generally, properties with large lots are country estates.

Looking at the variable *Build year* we notice that the maximum value is 2026, these are buildings under construction sold as project sales.

Considering the job specific variables, we notice that they are also right-skewed. The maximum job density is 1,432 and is found near Copenhagen. The properties farthest from an employment center are located in the municipality of Bornholm.

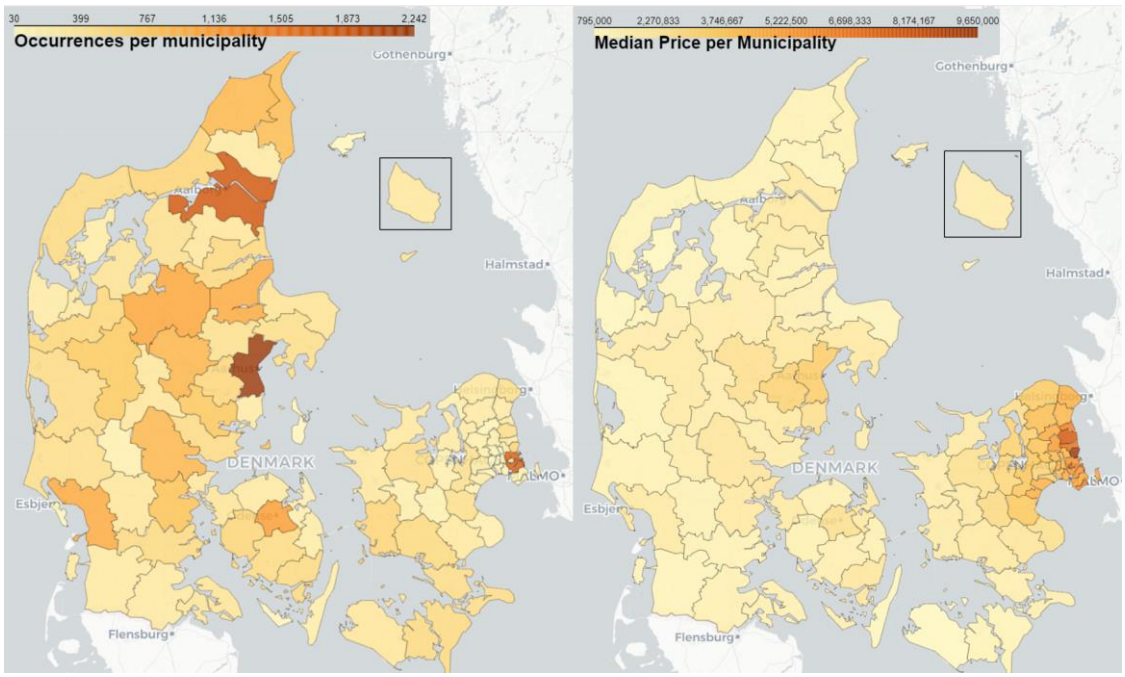
The infrastructure variables follow a similar pattern. Properties with 63 departures per hour are all located near Copenhagen Central Station, while those farthest from a station are in Bornholm.

5.2. Distribution of selected categorical property variables

Figures 5.1a and 5.1b present heatmaps showing real estate listings and median property prices across Danish municipalities, respectively.

Figure 5.1: Distribution of properties for sale and median price by municipality

(A) Distribution of properties for sale by municipality (B) Median price by municipality

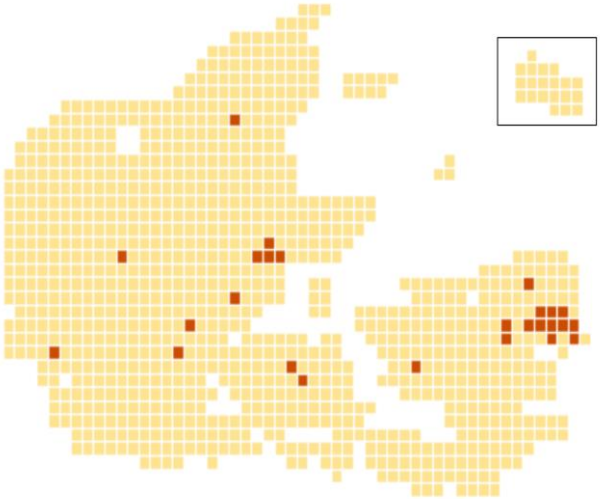


The data reveals that more properties are for sale in municipalities with larger cities, such as Copenhagen, Aarhus, and Aalborg, which aligns with their higher population densities. In Figure 5.1b, a distinct geographical distribution of real estate prices is evident, with the most expensive properties concentrated around Copenhagen.

5.3. Distribution of employment centers

To visualize the geographical distribution of the EC's identified in Section 4.3, Figure 5.2 displays a map of Denmark overlaid with 10x10 km grids. The red squares indicate the grids defined as EC. The figure shows that most ECs are concentrated around Copenhagen, with additional centers identified in Odense, Aarhus, Aalborg, and Esbjerg.

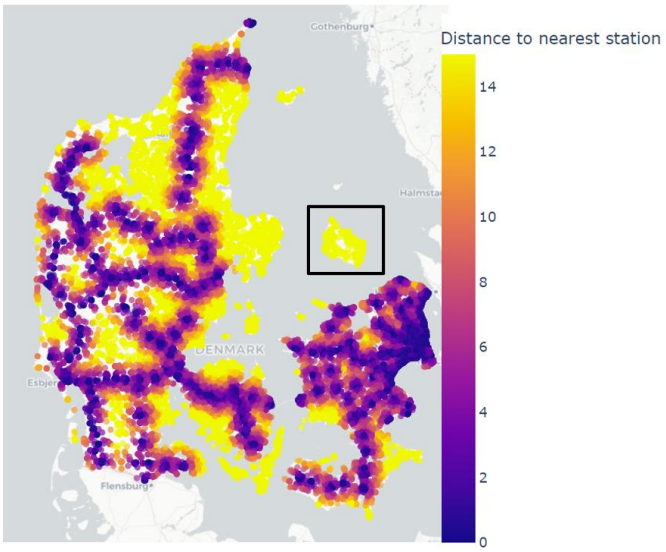
Figure 5.2: Illustration of 10x10 km grids and location employment centers



5.4. Distribution of train stations

Figure 5.3 displays all properties in our dataset along with their corresponding air route distances to the nearest train station.

Figure 5.3: Mapping of properties for sale and their corresponding distance to nearest train station

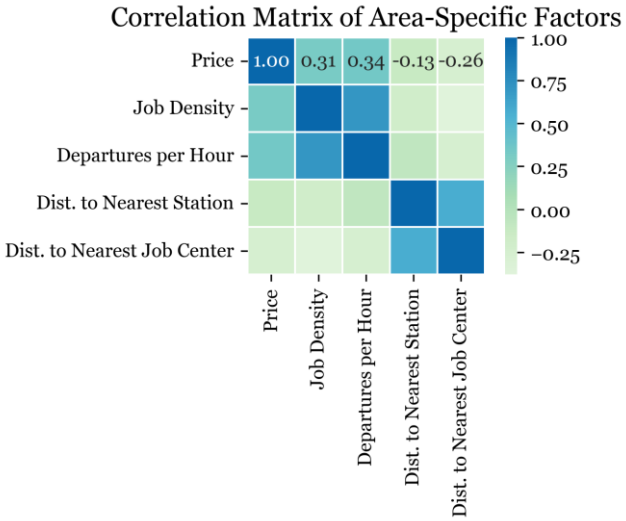


The map highlights the Danish railroad network, revealing a more developed rail infrastructure in Zealand, while the northern parts of Jutland are notably distant from train stations.

5.4. Correlation of measures

To obtain an idea of how features for employment and infrastructure correlates to real estate prices figure 5.3 shows the correlation between the price and our constructed variables.

Figure 5.3: Correlation between price and features of interest



The correlation matrix reveals a positive relationship between *price* and both *job density* and *departures per hour*. Whereas it shows a negative correlation between *price* and distance to EC's and nearest station. These relationships suggest that areas with higher job density and better public transport have higher property values. The correlation could support our hypothesis that area-specific characteristics enhancing professional opportunities significantly influence house prices, potentially improving our model's predictive accuracy.

6. Methods

To identify any complicated relationship between our input variables and the price of a property we use Machine Learning. Because we have a labeled dataset with both known input and output variables, we use supervised Machine Learning to train the model. Thus, we train our model by letting the algorithm compare its model-based outcomes with the true outcome given by the labeled training dataset and modify the model accordingly.

One possible downside of Machine Learning is the risk of overfitting the model to the training data. In this section we therefore describe the different methods we use in order to balance the trade-off between overfitting and underfitting our model. Since our outcome variable (*price*) is a continuous variable, we will only use regression techniques.

6.1. LASSO

LASSO is a regularization technique (L1) that helps to balance the bias-variance trade-off for linear regressions. It does this by introducing a penalty term to the usual minimization problem of a linear regression. The LASSO minimization problem can be written:

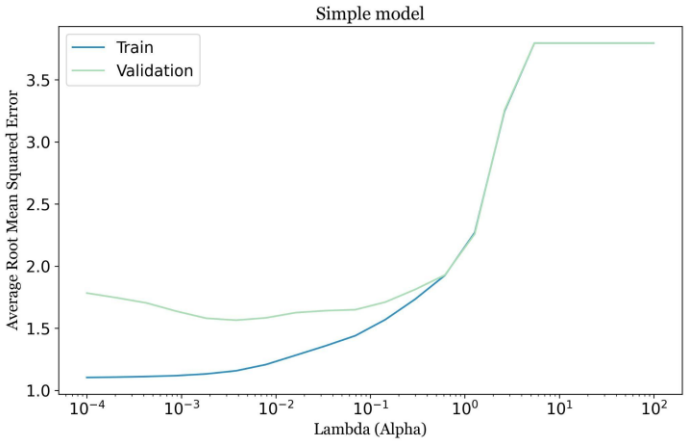
$$arg\ min_{\beta}\ \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2\ +\ \lambda\sum_{j=1}^{p}|\beta|$$

Where the last term is the penalty term. LASSO will then punish irrelevant features by setting the coefficients to zero. Thereby, LASSO decreases variance and mitigates the risk of overfitting the model. LASSO is especially useful to simplify the model.

In search of the optimal λ value for LASSO we fit the model using values of λ in the interval $[10^{-4}; 10^{2}]$ and we do 5-fold cross-validation as described in section 6.5. The optimal value of λ can be seen in Table A.1 in appendix A. Figure 6.1 shows the validation curve of the simple model. It shows that the AMSE for the validation data first decreases and then increases as λ increases. This happens because the model is overfitted to the training data for low values of

λ. Whereas the model is too general and underfitted to the data for higher values of λ. The validation curve for the extended model can be seen in Appendix A.2.

Figure 6.1: Validation curve for simple model using LASSO



6.2. Ridge

Ridge is another type of regularization (L2) used for linear regressions. The Ridge minimization problem can be written:

$$arg\ min_{\beta}\ \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + \lambda\sum_{j=1}^{p}\beta^2{}_j$$

The Ridge technique punishes irrelevant features by minimizing the coefficients towards zero. However, coefficients will never be set to zero, but just shrinked. This is also done to mitigate the risk of overfitting the model. Ridge is especially useful when dealing with multicollinearity and we use this regularization technique because we suspect multicollinearity.

In search of the optimal λ value for Ridge we fit the model using values of λ in the interval $[0; 10^{10}]$ and we do 5-fold cross-validation as described in section 6.5. The optimal value of λ can be seen in appendix A.1. Figure A2.2 in appendix A.2 shows the validation curve of both the simple and extended models. The intuition of the validation curve is similar to the one described in section 6.1.

6.3. ElasticNet

ElasticNet is a type of regularization that combines LASSO (L1) and Ridge (L2) regularization. The minimization problem of ElasticNet can be written:

$$arg\ min_{\beta}\ \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + \lambda_1 \sum_{j=1}^{p}|\beta| + \lambda_2 \sum_{j=1}^{p}\beta^2{}_j$$

By combining the penalties from LASSO and Ridge regularization ElasticNet allows us to deal with variable selection and multicollinearity at the same time. In addition to the λ-value the ElasticNet regularization also includes a L1 ratio measure that balances the impact of L1 and L2 regularization based on the data. In the extreme case where the L1 ratio is equal to 0 the Elastic Net regularization de facto becomes a ridge regression, whereas with a L1 ratio of 1 the ElasticNet regularization becomes a Lasso regression.

Because the ElasticNet has two hyperparameters the number of fits increase rapidly when one introduces more L1 ratios and λ values, because a fit has to be performed for each combination. We use the RandomizedSearchCV function from sklearn to preset the number of iterations we want to run. The RandomizedSearchCV function then randomly draws combinations of L1 ratios and λ-values from within defined boundaries.

We set no limit on the L1-ratio meaning that the random search function can draw ratios in the span of 0 to 1. The λ-values are based on some previous runs, which enabled us to narrow down on the most efficient λ-values. We draw 100 λ-values from a normal distribution with a mean value of 0.04 and standard deviation 0.02.

To find the optimal hyperparameters we used the RandomizedSearchCV function. We ran 35 iterations with 5 folds resulting in 175 fits for both the simple and the extended model (for chosen hyperparameters see appendix A.1). Using these hyperparameters we then refitted the model on the complete development data before predicting out of sample.

6.4. Random Forest

Like LASSO, Ridge and ElasticNet, Random Forest (RF) is a regularization technique for regression. RF differs from the other techniques by being an ensemble technique, meaning it combines multiple models instead of using a single model to reduce overfitting and improve

accuracy[4]. Specifically, RF uses an ensemble of Decision Trees (hence the name) and combines the predictions made by each tree in the forest.

Random Forest enables us to capture complex non-linear interactions between features, since it does not assume any particular form of relationship between our features and target variables. We therefore believe that it can be a valid method for accurate predictions in the real estate market, in which multiple variables are assumed to interact - sometimes irregularly due to geographical and demographic differences.

ML Implementation

We used a Random Forest model with a randomized grid search and cross-validation. The search space for hyperparameters is defined as follows:

We choose the parameter space with the following considerations:

- Number of Trees (n_estimators): Balances model complexity with computational efficiency, ranging between 100 and 500 trees.
- Maximum Depth of Trees (max_depth): Controls the complexity of the patterns the model can capture, with values between 10 and 100, helping to balance complexity and overfitting.
- Minimum Samples to Split a Node (min_samples_split**)**: Set between 2 and 20, this parameter controls the trade-off between underfitting and overfitting.
- Minimum Samples at a Leaf Node (min_samples_leaf): Ranges from 1 to 10, ensuring more balanced trees and helping to prevent overfitting.
- Maximum Features for Splitting (max_features): Allows for 10% to 90% of features to be considered at each split, balancing model complexity with feature randomness.

The grid configuration allows us to search a wide range of model configurations. The randomized grid speaks to it remaining computationally feasible.

6.5. Cross-Validation

To further mitigate the risk of overfitting we use k-fold cross-validation. That is, after splitting the data set in a development (66%) and test (33%) data set we further split the development data in k folds. We then leave one fold out to validate the model on and train the model on the remaining k-1 folds. We repeat this process k times, leaving out a new fold each time. Finally, we chose the hyper parameters that on average gives the lowest mean-squared-error

[4] Random Forest Regression in Python - GeeksforGeeks

and we then train the model on the whole development data set. For all regularization techniques we set the number of folds to 5.

7. Results

In this section we present our results using the 4 methods described in section 6 and 2 different models: (i) A simple model using only property specific features to predict our outcome variables, and (ii) an extended model using property specific and area specific features. When using the regularization techniques of LASSO, Ridge and ElasticNet we use a linear regression with a polynomial degree of two.

Table 7.1 shows our results using the optimal hyperparameters (for hyperparameters see appendix A1).

Table 7.1: Results using different regularization techniques

	ML Features			**LASSO**		**Ridge**		**ElasticNet**		**Random Forest**	
	Property specific	Job related	Mobility related	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Simple	✓	-	-	2.16	0.68	3.153	0.421	2.141	0.634	1.511	0.828
Extended	✓	✓	✓	1.75	0.756	3.287	0.751	1.684	0.773	1.477	0.826

Table 7.1 shows that, comparing a simple model, and our extended model, we find that the extended model performs the best for all models. Only in the Ridge setting are we not able to improve the R^2, we are however able to improve the RMSE.

Our results indicate that adding area specific features concerning job opportunities and railway infrastructure improves our prediction of real estate prices. The overall best model is the RF extended model. The following therefore focuses on the results from RF extended models. Results from the remaining models can be found in Appendix A.2 and A.3.

Figure 7.1 compares RF model prediction to actual values in the test data for the extended model to those of the simple model. The figure shows that both models perform worse on more expensive properties - whereas the absolute error is less for less expensive properties. This is expected as we have a few extreme outliers with very high prices in our data, hence our models have limited data points to train and learn from in these cases. Furthermore, the

price on very expensive properties might well be determined by other factors than what appears in our data.

Figure 7.1: Comparison of model predictions vs. actual using the extended RF model

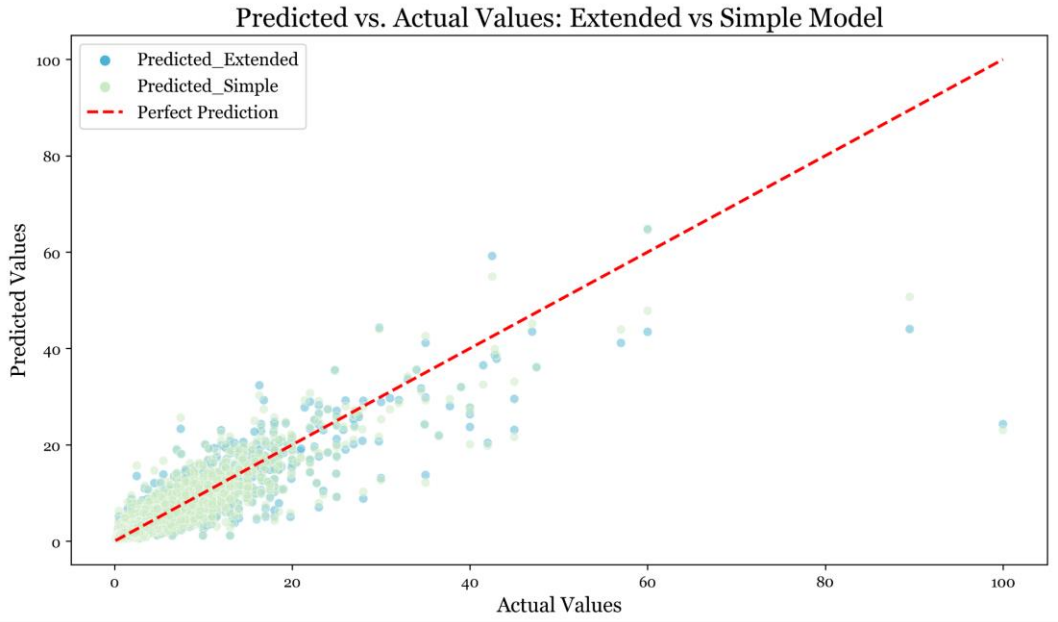


Figure A3.1-A3.4 in appendix A.3 show the same result for the remaining models.

Table 7.2 shows the most important features using the RF extended model. The feature *distance to EC* is in the top 5 most important features. Appendix A.4 shows similar results for the remaining extended models which also includes departures per hour in the most important features. This indicates that in fact distance to professional opportunities and railway infrastructure can improve our predictions of real estate prices.

Table 7.2: Most important features using extended RF model

	Features
1	Basement Size
2	# Rooms
3	Energy Class D
4	Distance to Employment Center
5	Municipality == 329

8. Discussion

All four ML models tend to over-predict low values and under-predict higher ones, likely due to extreme outliers on the target variable price and features. An ability to handle outliers could well have proved crucial for model selection. We end up preferring a RF model. RF's are known to have both strengths and weaknesses.

We also found that our focus on 'distance to nearest station' made our model very sensitive to including/excluding listings on Bornholm, which doesn't have rail connection. All models, especially Ridge, perform better without Bornholm.

Future research should consider excluding farms from the dataset. Agricultural properties, with their larger lot sizes and greater distances from employment centers and stations, likely have different valuation drivers compared to residential properties like villas, terraced houses, and flats.

One could make a similar argument for disassociating villas and flats. Perhaps these are determined by demography? One gets a sense of this by considering the expected sign of 'Distance to Employment Center'. This is far from straightforward to predict. Younger, childless workers often prefer to live close to urban centers, where they can be within walking distance of their workplaces. In contrast, parents with young children may prioritize access to quality schools, which is often placed within suburban single-family neighborhoods (Kim and Jin, 2019; Schimer et al., 2014). In our model, where we lump everything together as "real estate" we are perhaps asking too much?

Our conception of real estate assumes a certain degree of homogeneity amongst property buyers. Malpezzi (2003) and Zietz et al (2007) argue that real estate buyers are heterogenous. Simply put, different socioeconomic groups value things differently. One's preferences for specific home features, like size and amenities, tends to shift depending on what price segment of the housing market one belongs to. Adding demographic features to our model could help account for this.

The feature 'distance to Employment Center assumes a centre. Other urban development scholars have argued that polycentric models are more appropriate (see, for instance, Anas & Richard, 1998; Giuliano & Small, 1991; McMillen, 2001). They find that, as metropolitan areas expand and evolve, the trend is often towards a *constellation* of employment sub-centres, which challenges the notion of a single, dominant business district.

Another clear limitation of our study - and one which prevents it from being directly replicable - is that we end up using a somewhat ad hoc measure for Employment Centers. We find that Giuliano and Small's standard definition of ECs, outlined in the Literature Review, is too stringent for our dataset. We instead opt for a relative definition, and settle on a quantile, which creates a total of 25 employment centers. What this achieves for us is to create a measure that is spread out across Danish geography. This permits us to apply our analysis to real estate listings throughout the country. It has the downside that the measure is unlikely to be directly transferable to other local/national contexts. Defined only in terms of 'number' of jobs, it also lacks a bit of nuance. Future studies could well augment the measure to capture what type of jobs.

Finally, our model does not account for spatial autocorrelation, which likely exists in housing price data. Nearby properties tend to have similar characteristics and prices, violating the assumption of independence between observations. The use of standard k-fold cross-validation with spatially autocorrelated data may cause information leakage between nearby locations in training and test sets. This could result in poor generalization. Future studies would benefit from implementing spatial cross-validation techniques and considering spatial lag models to address this limitation.

9. Conclusion

In conclusion, this study successfully integrated data from real estate listings, job postings, and train station schedules to create a comprehensive dataset capturing various factors influencing real estate values. By applying four machine learning models—LASSO, Ridge, Elastic-Net, and Random Forest—we compared the predictive power of a simple model based solely on real estate features with an extended model that included additional variables such as proximity to job centers, distance to the nearest train station, and train departures per hour. The results consistently demonstrate that the extended model outperforms the simple model across all cases, highlighting the importance of incorporating a broader range of contextual features when modeling real estate prices. These findings underscore the value of integrating diverse datasets for improved accuracy in real estate prediction models.

Appendix

A.1 Hyper Parameters

Table A.1: Optimal hyperparameters

		LASSO	Ridge	Elastic-Net	Random Forest
Simple	lambda1	0,0038	-	0,01126	-
	lambda2	-	316		-
	L1 ratio	-	-	0,2121	-
	n_estimators	-	-	-	335
	max_depth	-	-	-	82
	min_samples_leaf	-	-	-	1
	min_samples_split	-	-	-	8
	max_features	-	-	-	0.63
Extended	lambda1	0,0038	-	0,01199	-
	lambda2	-	825		-
	L1 ratio	-	-	0,8585	-
	n_estimators	-	-	-	202
	max_depth	-	-	-	16
	min_samples_leaf	-	-	-	1
	min_samples_split	-	-	-	4
	max_features	-	-	-	0.6

A.2 Validation Curves

Figure A2.1: Validation curves for the simple and extended LASSO

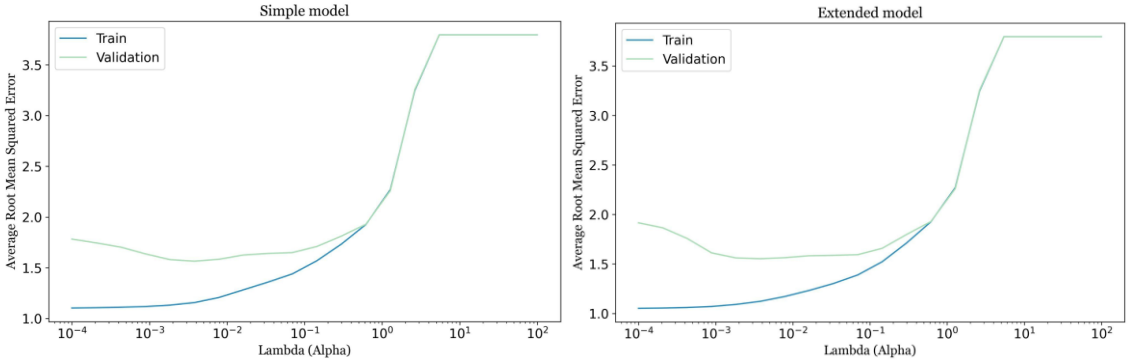


Figure A2.2: Validation curves for the simple and extended Ridge

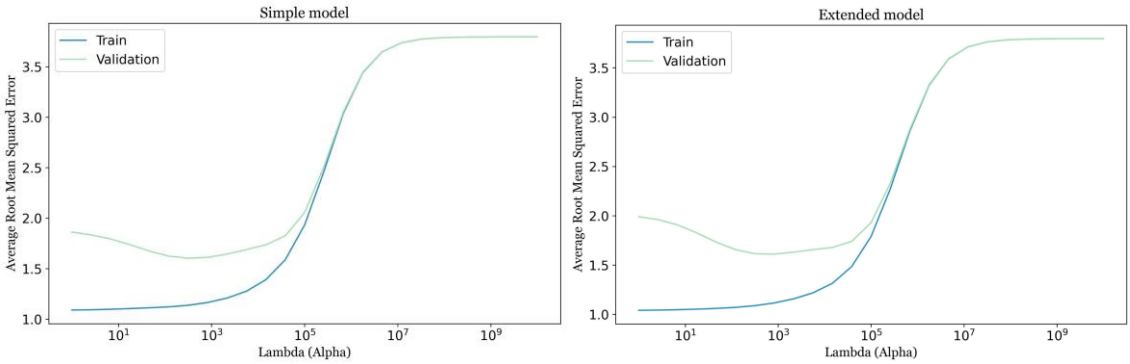


Figure A2.3: Validation plots for the simple and extended ElasticNet

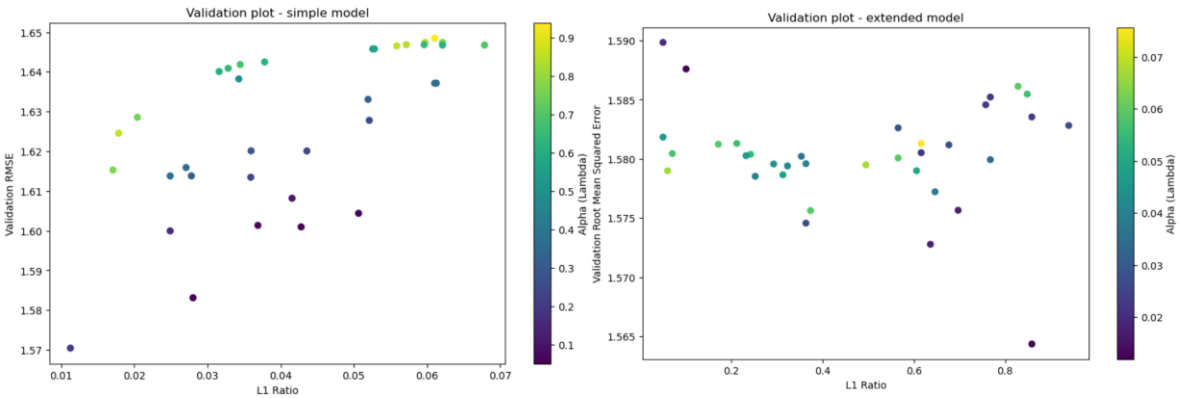
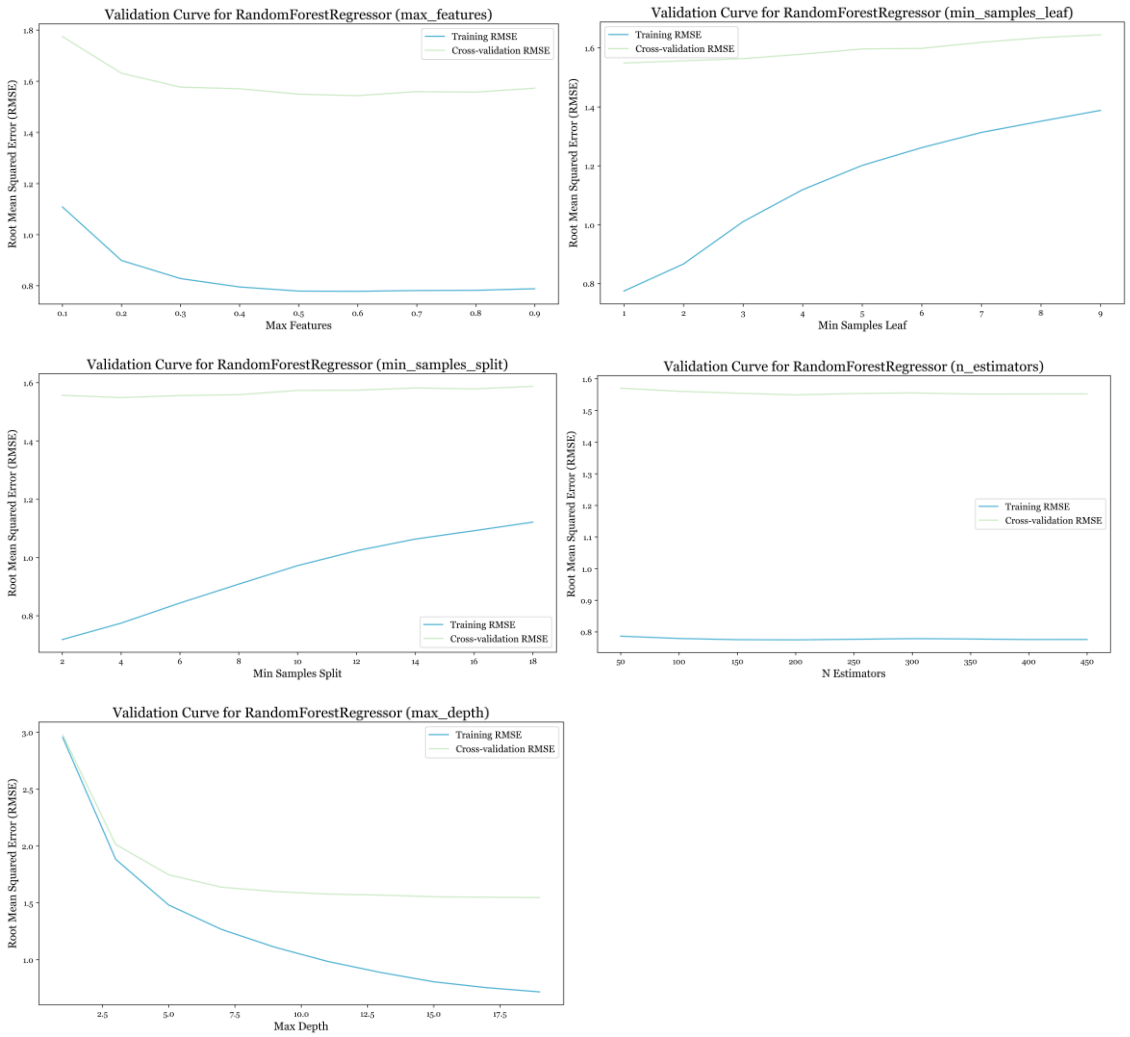


Figure A2.4: Validation curves for the extended Random Forest



A.3 Predicted vs. actual values

Figure A3.1: Predicted vs. actual values the simple and extended LASSO

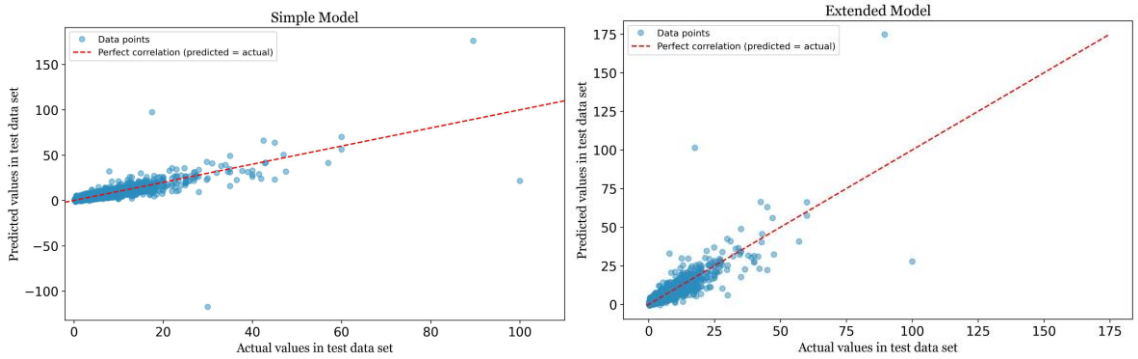


Figure A3.2: Predicted vs. actual values the simple and extended Ridge

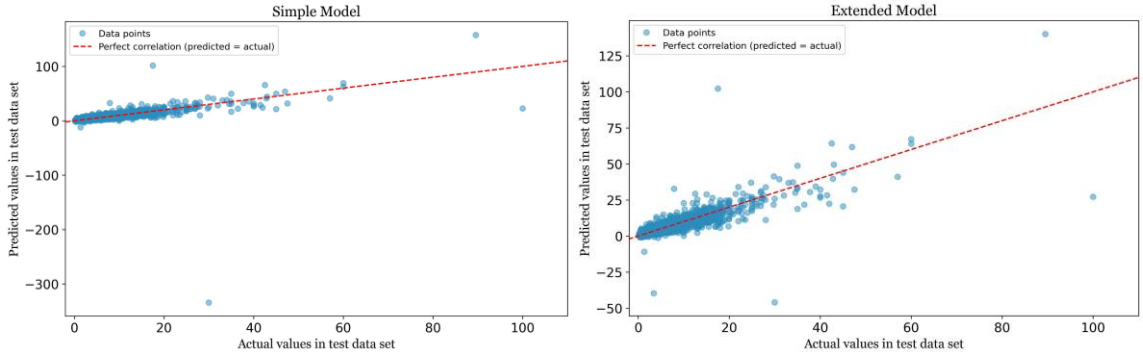


Figure A3.1: Predicted vs. actual values the simple and extended ElasticNet

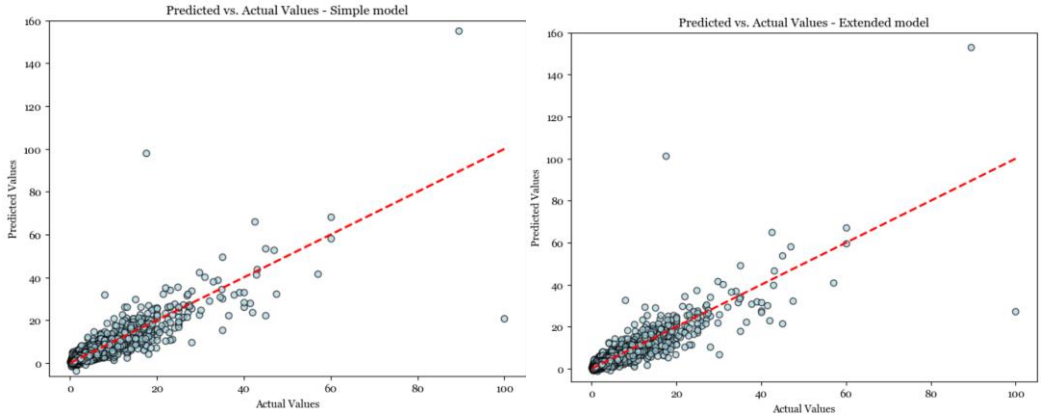
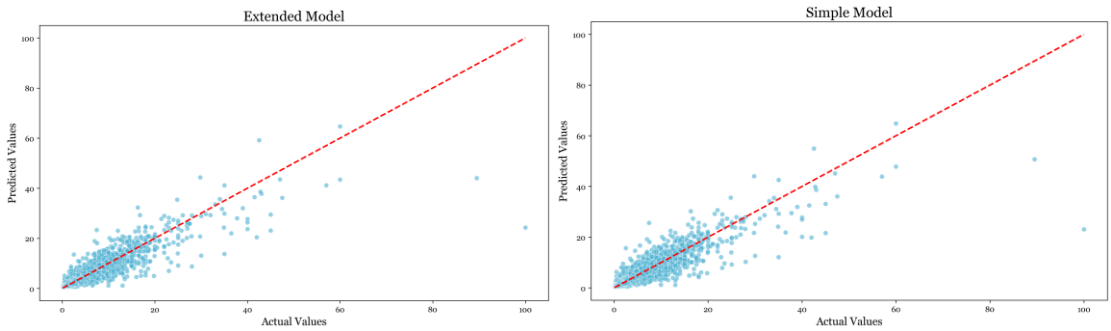


Figure A3.1: Predicted vs. actual values the simple and extended Random Forest



A.4 Most important features

Table A4.1 Top5 important features in extended LASSO

	Top 5 Features with positive Sign	Top 5 Features with negative Sign
1	Monthly expenses	Monthly expenses X Apartment
2	Size X Lot size	Monthly expenses X municipality == 846
3	Job density X Lot size	Rooms X Lot size
4	Departures per hour X monthly expenses	Job density X monthly expenses
5	size X Apartment	Monthly expenses X Agricultural property

Table A4.2 Top5 important features in extended Ridge

	Top 5 Features with positive Sign	Top 5 Features with negative Sign
1	Monthly expenses	Monthly expenses X Apartment
2	Monthly expenses X build year	Monthly expenses X municipality == 846
3	# Rooms X monthly expenses	Departures per hour X municipality == 157
4	Size X lot size	Monthly expenses X Agricultural Property
5	Monthly expenses X Energy class C	Basement size X Apartment

Table A4.3 Top5 important features in extended ElasticNet

	Top 5 Features with positive Sign	Top 5 Features with negative Sign
1	Monthly expenses	Monthly expenses X Apartment
2	Monthly expenses X build year	Monthly expenses X municipality == 846(Mariagerfjord)
3	Size X lot size	Monthly expenses X Agricultural property
4	Rooms X Monthly expenses	Monthly expenses X municipality ==223(Hørsholm)
5	Size X apartment	Monthly expenses X municipality ==173((Lyngby-Taarbæk)

Bibliography

Agnew, K. & Lyons, R. (2018). The impact of employment on housing prices: Detailed evidence from FDI in Ireland. *Regional Science and Urban Economics*.

Anas, A.; Richard, A.; Small, K. Urban spatial structure. *J. Econ. Lit.* 1998, *36*, 1426–1464.

Andersen, L. (2021). *Boligmarkedet skaber kløfter. Det er en ulighedsgenerator*. [online] Available at: https://www.ae.dk/node/2834/pdf-export [Accessed 27 Aug 2024].

Armstrong, R. & Rodriguez, D. (2006). An evaluation of the accessibility benefits of commuter rail in Eastern Massachusetts using spatial hedonic price functions. *Transportation*, 33, 21–43.

Gibbons, S. & Machin, S. (2005). Valuing rail access using transport innovations. *Journal of Urban Economics*, 57(1), 148–169.

Giuliano, G. & Redfearn, C. (2005). Employment centers and residential location: A case study of Los Angeles. *Urban Studies*, 42(7), 1221-1240.

Giuliano, G. & Small, K. (1991). Subcenters in the Los Angeles region. *Regional Science and Urban Economics*, 21(2), 163-182.

Hess, D. B. & Almeida, T. M. (2007). Impact of Proximity to Light Rail Rapid Transit on Station-area Property Values in Buffalo, New York. *Urban Studies*, 44(5-6), 1041-1068.

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111.

Kim, D. & Jin, J. (2019). The Effect of Land Use on Housing Price and Rent: Empirical Evidence of Job Accessibility and Mixed Land Use. *Sustainability*, 11(3), 938.
McMillen, D. P. (2001). Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3), 448-473.

Zhou, Z., Chen, H., Han, L. & Zhang, A. (2021). The Effect of a Subway on House Prices: Evidence from Shanghai. *Real Estate Economics*, 49(2), 199-234. https://doi.org/10.1111/1540-6229.12275