KØBENHAVNS UNIVERSITET
ØKONOMISK INSTITUT

Predicting Valuation Prices of Danish Real Estate Property

- Undertitel

Group 40.

Exam numbers: 115, 144, 146, 204.

Contribution:

115:

144:

146:

204:

Afleveret den: 30/08/2019

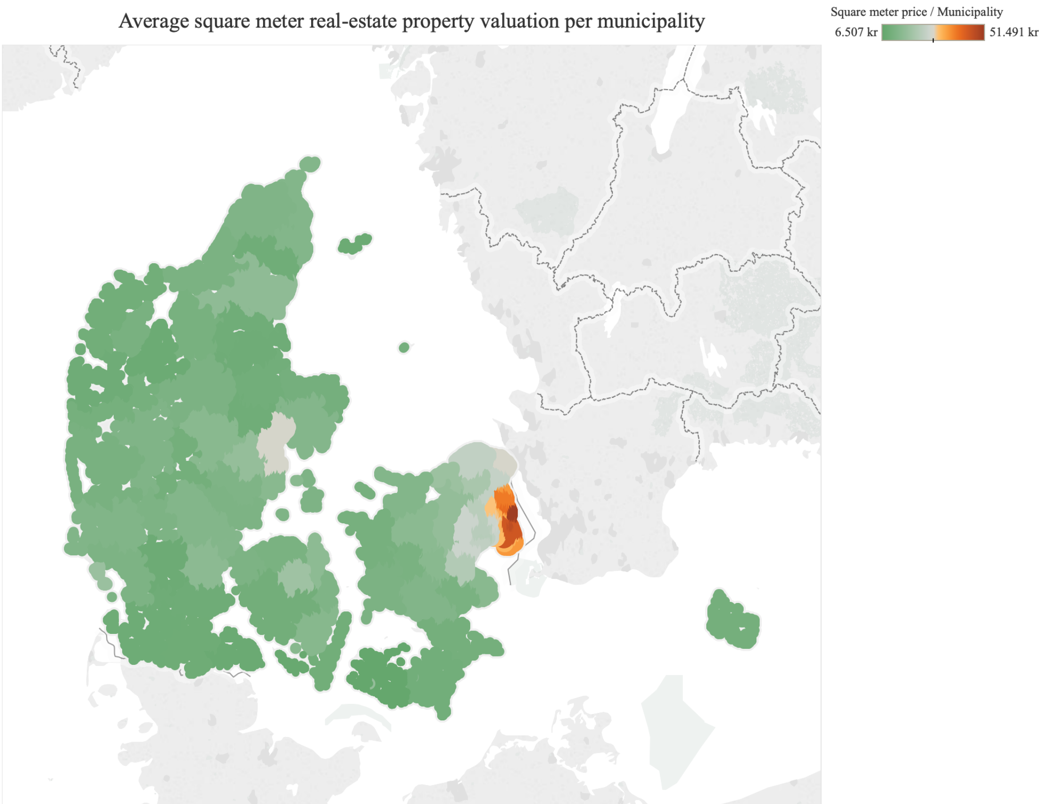
Typeenheder:

Contents

1 Introduction

This research paper regards listed Danish real estate property valuations. Since 2011, the annual number of real estate properties sold has slightly increased, with 2017 being the year with most sold properties in 10 years (Danmarks Statistik, 2018). The price of real estate properties is also rising through 2019 (Danmarks Statistik, 2019). Although this generally depicts a willingness of buyers to pay more for real estate property, and higher valuations from realtors, a closer analysis of active valuations adds a local view of which conditions contribute to the valuation of a real estate property.

Figure 1



Source: Own creation, with data from Boliga.dk

Figure 1 shows the average square meter price valuation of active offers per municipality in Denmark. The maximum square meter price being approximately 8 times higher in the municipality of Gentofte, than Lolland, the municipality with the lowest average square meter price valuations. A tendency visualised in the figure, is that average square meter price valuations are higher in highly populated municipalities and in suburban areas surrounding Copenhagen.

The valuations included in the research are collected from one of Denmarks largest online real estate websites named Boliga.dk. The Boliga data contains approximately 66,000 active offerings with valuations ranging from DKK 15,000 to DKK 85 million. This research paper intends to use machine learning to analyse real estate price valuations and which geographical and sociodemographic criteria affect valuations. Our research question is as follows:

How can geographic and sociodemographic features be utilised to predict real estate property valuations in Denmark?

This research paper contains a section describing the construction of research data and assesses the choices made in gathering meaningful features for the dataset. Also, a section regarding the choice and optimisation of machine learning models is included, where the intention is to provide insights on our progress of finding the optimal model. As a result, the best performing model is chosen with a discussion of its usability.

2 Literature Review

Within social data science there exist different definitions of big data and machine learning. This section seeks to clarify the use of these terms within the scope of this paper.

2.1 On Big Data & Machine Learning

Historically, big data has been a term reserved for data that was unable to be processed by extant software. However, recent increases in computational power has enabled data processing of hitherto unheard data quantities[1]. Today, big data is no longer too cumbersome to analyse. Matthew Salganik[2] instead mentions ten typical characteristics of big data. Salganik's tentative definition suggests that big data is not a single entity, but includes many different types of systems. Among the most important features of Salganik's definition of big data is that the data has a high frequency of observations and is continuously being generated. Since the data is always-on it is also drifting, meaning that the structure and population it represents is ever-changing. It is therefore important for scientist to understand that big data is not a naturally occurring system, but driven by the engineered purpose of the system. This algorithmic confounding force the scientist to be careful regarding any observed human behaviour that is extracted from a single digital system.
In the analysis of big data it is often useful to employ machine learning as it can identify and predict non-linear relationships in big data sets, that otherwise would remain hidden. Data scientists' have for the most part optimised the predictive capabilities of the algorithm applied, and as a consequence they have often ignored or trivialised machine learning's potential in causal modelling[3].

[1]Lazer, David and Jason Radford (2017) *Data ex Machina. Introduction to Big Data*. Annual Review of Sociology

[2]Salganik, Matthew J. (2018) *Bit by Bit Bit - Social research in the digital age.*

[3]Varian. Hal R. *Big data: New tricks for econometrics*

The technical and theoretical challenges faced by big data and machine learning research are important to consider, when employing the tools they provide. One of the major discussions revolve around machine learning's predictive capabilities. Chris Anderson[4] argues that since the computing power and the scale of data has increased exponentially, our reliance on scientific models could become obsolete. Instead of focusing on the theoretical implications of observations, scientists should, according to Anderson, focus on the statistical outputs: In the age of big data, correlation perhaps should supersede causality and consequently social data scientists should not try to develop coherent models or unified theories to explain a social phenomenon.
However, many social data scientists argue against this point of view. Justin Grimmer[5] claims, that correlations extracted from big data cannot stand alone. Large quantities of data is not sufficient to make scientifically valid causal inferences. It requires a rigorous research design and clear theoretical assumptions, in order to yield scientifically accurate estimates. Indeed the social sciences greatest contribution to big data research, comes from the organised framework provided by rigorously tested theory[6].
The contribution from the social sciences to machine learning help create new methods that will be able to utilise the strengths of machine learning to help solve causal inference problems within the framework of a well defined theory[7]. These new approaches could help define what variables to manipulate and how to properly use machine learning within the framework of theoretical assumptions. Ultimately, big data and machine learning could increase the scope of the social scientist's field, not only by delivering new data and methods, but by helping the social scientist to focus on new questions and avenues of research[8].

2.2 Ethical Concerns in a Petabyte Age

The ethical principles of social research are anchored in the fundamental human rights, which are broadly formulated in the UN Declaration of Human Rights. Additional policies and declarations that codify principles of research ethics and the ethical treatment of research participants include the Nuremberg Code, the Helsinki Declaration, the Belmont Report, and the Menlo Report[9] [10]. These codes and addendums originate mostly in the biomedical field, though they encompass the central principles applied to all human research, which have led some academics to call for a Hippocratic Oath for data scientists to safeguard against powerful new technologies

[4] Anderson, Chris. (2008), *The end of theory: The data deluge makes the scientific method obsolete*

[5] Grimmer, Justin (2015) *We are all social scientists now: how big data, machine learning, and causal inference work together*

[6] Einav, Liran and Jonathan Levin (2014) *Economics in the age of big data*

[7] Athey, Susan (2018) 'The Impact of Machine Learning on Economics' in Ajay Agrawal et al. (eds) *The Economics of Artificial Intelligence: An Agenda*

[8] Mullainathan, Sendhil, and Jann Spiess (2017) *Machine Learning: An Applied Econometric Approach*

[9] Salganik, Matthew J. (2018) *Bit by Bit Bit - Social research in the digital age.*

[10] European Commission (2018). *Ethics in Social Sciences and the Humanities*

under development in laboratories and tech firms[11] [12]. This discussion is nothing new however, as a tentative reformulation of the Hippocratic Oath was introduced by Karl Popper[13], wherein he stressed the importance of professional responsibility, a critical mind, and an overriding loyalty towards the betterment of mankind.

Matthew Salganik offers four principles deduced from the Belmont and Menlo Report that should guide the ethical deliberations of the researcher: 1) the respect for persons, that is individuals should be treated as autonomous and if circumstances require it individuals should be entitled to additional protections. 2) Beneficence stresses the importance of doing no harm and to maximise the possible benefits and minimising any potential harms. 3) The principle of justice touches upon the importance of the distribution of burdens and benefits of the social scientist's research. This principle stresses that is should not be a single stratum of society that bears the costs of the research while another stratum benefits. 4) The fourth and final principle is the respect for law and public interest, according to Salganik, the principle consists of two distinct elements, that is compliance to relevant laws and legal contracts and transparency-based accountability. It is worth noting that Popper's tentative Hippocratic Oath mirrors the first three principles put forth by Salganik, stressing the importance of the ethical conduct of the researcher.

The need for rigid ethical standards within computational social science was made apparent by the Cambridge Analytica Scandal that broke on the 17^{th} of March 2018. Where Steve Bannon could reveal that between 2013 and 2015 Cambridge Analytica had exploited a loophole in Facebook's API. This allowed the company to harvest profile data from 87 million Facebook users, without the user's permission and used the harvested data to construct a massive targeted marketing database based on the user's likes and interests[14]. Other examples of misuse of data acquired from Facebook include the Harvard-run experiment, where students' data was used to create new knowledge about how social networks form. The gathered information was used to examine how these networks and their actors' behavior co-evolve. Another example of misuge of data is the emotional contagion experiment from 2012, where approximately 700,000 Facebook users were involved in a research experiment to examine the extent to which a person's emotions are affected by the emotions of the people they interact with (Salganik. 2018).

From this it should be evident that clear ethical guidelines are required in order to protect the user's privacy from tech-savvy companies. To this end, the European Parliament introduced the General Data Protection Regulation (GDPR). With its seven overarching principles the GDPR seeks to formalise the procedures involved in the data processing and storing of sensitive and private information (CDRC[15]). These principles include and expand upon the principles found

[11] Rotblat, Joseph (1999) *A Hippocratic Oath for Scientists*

[12] Sample, Ian (2019, Fri 16 Aug 2019) *and tech specialists need Hippocratic Oath, says academic.*

[13] Popper, Karl (1969) *The Moral Responsibility of the Scientist*

[14] Vox.com (2018) *The Cambridge Analytica Facebook scandal* [Online]

[15] Consumer Data Research Center, UK [CDRC] (2018) *The General Data Protection Regulation & Social Science Research* [Online]:

in the Belmont and Menlo Report. The most important consideration, however, must be that even a dataset comprising tens of thousands observations involve human beings who must be protected from adverse side-effects of the social research. There is considerable evidence that points to the fact that even in anonymised data sets it can be possible to backtrack an individual's identity. A researcher must therefore be mindful of the mosaic effect if the dataset combines large amount of data from various sources (European Commission 2018).

3 Data Description & Ethics

In the following sections we will consider the ethical ramifications of our inquiry. We will focus our efforts on the contractual obligations laid down by Boliga's Terms & Conditions and the more general ethical considerations a social scientist should consider before conducting research. Following this, will be a brief description of our data, how it was collected, cleaned and examined.

3.1 Ethical Considerations in the Current Research Project

In the current paper the appropriate care and consideration has been given to the ethical concerns regarding the scraping, processing, and the presentation of the data. Drawing upon the European Comission's[16] guidelines and principles for ethical conduct in social data science, the potential harm to users of Boliga's website were carefully considered. As a step to prevent the mosaic effect and in an effort to anonymise the scraped data, only aggregated data will be presented in this paper, so that no single observation can be identified from the analysed data.
Informed consent has not been obtained from the users of the site, prompting us to consider the consequences of the lack thereof, as informed consent is paramount to the proper, ethical conduct in social science. However, as in this instance, informed consent can be logistically impossible to collect from all participants in the study. Salganik mentions that informed consent for everything is an ideal, but in practice impossible to obtain and researchers should instead strive to follow an alternative rule, that he describes as: "some form of consent for most things." [17] Adhering to this more complex understanding of the practicality of informed consent, we chose to contact Boliga to inform them of our intent to scrape their website and use the data in an educational context[18]. Boliga responded positively to our inquiry, which we took as informed consent from a third party on behalf of the participating users in our study.
In considering the legal ramifications of our research and to make sure we adhere to the seven principles of GDPR, and other appropriate legislation and legal contracts, we consulted the general guidelines introduced by the Consumer Data Research Center. In particular we noted that

[16] European Commission (2018). *Ethics in Social Sciences and the Humanities*
[17] Salganik, Matthew J. (2018) *Bit by Bit - Social research in the digital age.* p.303
[18] Shiab, Nael (2015) *On the Ethics of Web Scraping and Data Journalism*

we are justified to collect and use the data on the lawful basis of legitimate interests. Furthermore, we consulted with Boliga's Terms and Conditions to avoid any legal ramifications and the appropriate contractual terms of interest can be seen in §10 Terms and Conditions. In order to comply with these terms we refrained from burdening their website's performance by implementing a time.sleep function. This causes each scraping iteration to pause for 0.5 seconds before commencing on scraping the next page (see Jupyter).

3.2 Data Scraping Process

In the following paragraphs our scraping efforts will be described. The scrapers can be examined in the attached Jupyter Notebooks.

Boliga

Our data comes from Boliga.dk, the largest independent online web-portal for real estate sales in Denmark and has access to unique features such as days-for-sale, price-development and access to BBR - the Danish Building and Housing Register [19]. This gives unique insight into the pricing of real estate in all 98 municipalities of Denmark. 65,950 properties were for sale at the time of scraping.

The scraping process was conducted on Friday the 23^{rd} of August 2019. In order to scrape the data of interest we familiarised ourselves with the network-structure of Boliga. On the basis of these insights we constructed a code which was able to scrape every page, containing information on the currently listed real estates on Boliga. The scraper requested all information available from each individual page, which surmounted to 1,319 URL requests.

For each listed property, 34 features and the target variable (price) was collected. Appendix XX provides an overview of the features with a short description, and whether the feature has been dropped or saved for later usage. There are three reasons for a feature to be dropped:

1. The feature does not act with independent characteristics according to the research,
2. The feature contains insufficient data,
3. The feature is poorly formatted and cannot efficiently be recreated.

10 features are considered to be of interest.

Features Obtained:

Continuous:	basementSize, buildYear, ownersExspenses, lotSize , price, rooms, size
Categorical:	ForClosure, Type, Municipality

hvorlangterder.dk

hvorlangterder.dk returns distances from a given address, to everyday commodities such as supermarkets, hospitals and schools. The scraping of hvorlangterder.dk was achieved by writing a function that took in the GPS-coordinates gathered from Boliga. The function returned 65,950

[19] www.boligagruppen.dk

URL responses, from which relevant information was extracted.
As Jupyter performs poorly when running long asynchronous tasks and the estimated running time was 18 hours, this procedure was run in Visual Studio Code.

Features obtained:

Distances to: lake, forest, doctor, supermarket, school, daycare, hospital, train
metro, pharmacy, library, coast, junction

DAWA

Denmark's Addresses Web API[20] (DAWA) is an API containing data regarding Danish addresses. DAWA will be used to collect information about the municipality that each address is located in. This data will be used to merge the socio-demographic data collected from various sources (see below) with the data scraped from Boliga.

Sociodemographic Factors

Sociodemographic factors on municipality level was collected from statistik.politi.dk and Danmarks Statistik. The factors include income, reported crime, level of highest completed education, etc. These were transformed into ratios, by taking the total population in a given municipality into account.

Features obtained: unemployment_relative, primary_school_educ, high_school_educ
(relative to population) vocational_educ, SHE, MHE, bachelors_degree
LHE, avg_municipal_income_2017, Total_reported_crime
Population_in_urban_development , Socioeconomic_index
average_class_size, expenses_sport_and_other_cultural_activities
expenses_per_school_student

All educational features are a measure of highest completed education. Furthermore SHE, MHE, LHE are abbreviations of short-, medium- and long-cycle higher education.

3.3 Analysis of Scraping Logs

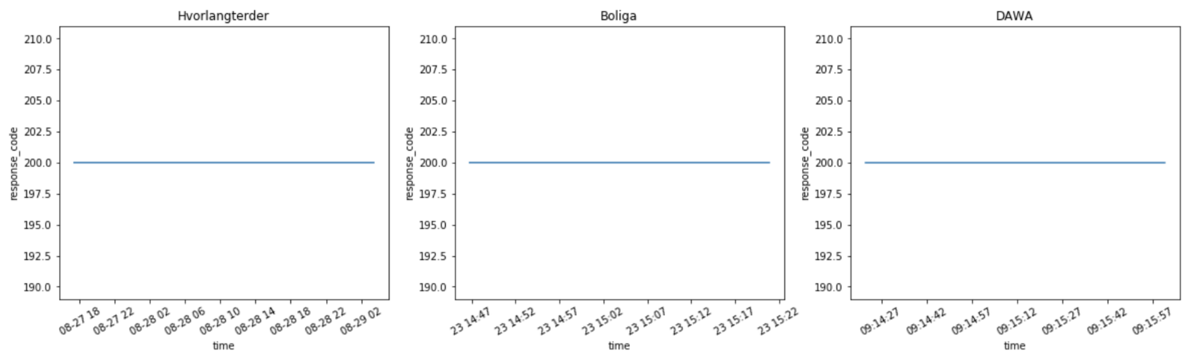
Three datasets were obtained through web scraping from Boliga.dk, hvorlangterder.dk and DAWA.dk. This section will analyse the logs created in the scraping process, with the intention of ensuring the data quality.

Response Code

When parsing data from websites, the goal is to receive a response code '200', expressing a successful request. Figure 2 displays response codes for the duration of the web scrapings. Every request received a response code 200, making further evaluation of the response codes redundant.

[20]red. Danmarks Adressers Web API

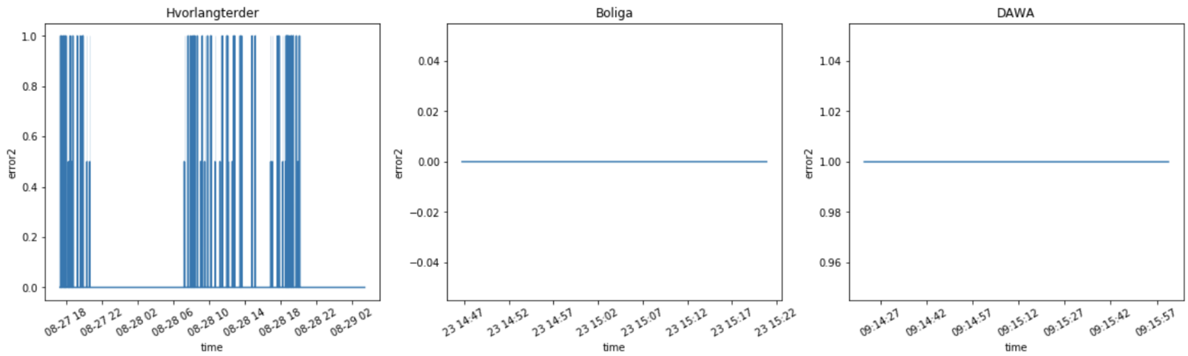
Figure 2



Error Codes

Figure 3 plots error codes received throughout the duration of the scrapings. Boliga and DAWA received no errors, while hvorlangterder received a total of 80 errors. The error code returned the same response in all 80 cases: **('Connection aborted.', RemoteDisconnected('Remote end closed connection without response'))**

Figure 3

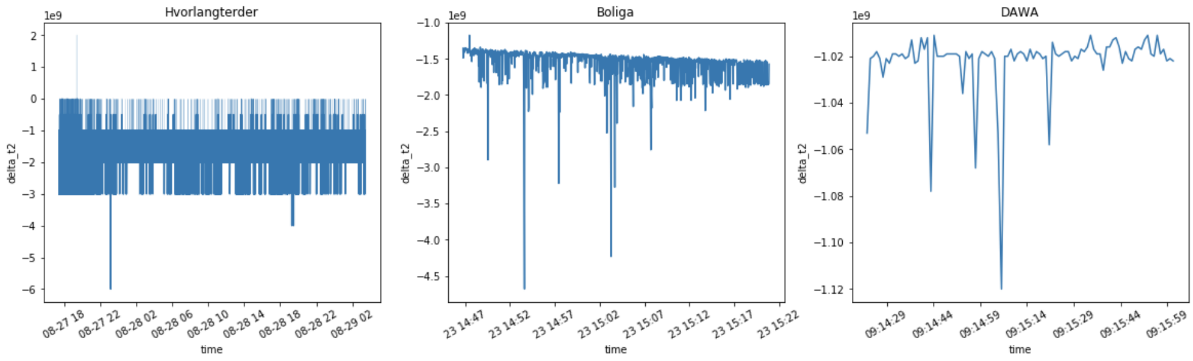


To understand the errors, a closer manual analysis of the log was performed on requests receiving errors. The errors occurred in random intervals and were subsequently followed by another request to an identical URL. Therefore, we assessed that these errors were due to connection problems or server traffic. Since the errors were corrected by a following request, it should not affect the research.

Server Response Times

An indicator of poor data quality is varying response times from the server being scraped. Figure 4 plots changes in response times for the duration of the web scrapings. The three data sources all resulted in few variations of response time. Unusual response times have been assessed, to be caused by a slow server connection, or due to larger quantities of data being requested. To validate the quality of data, manual control of the returned results and the website results was performed, without notice of significant corruptions in the data.

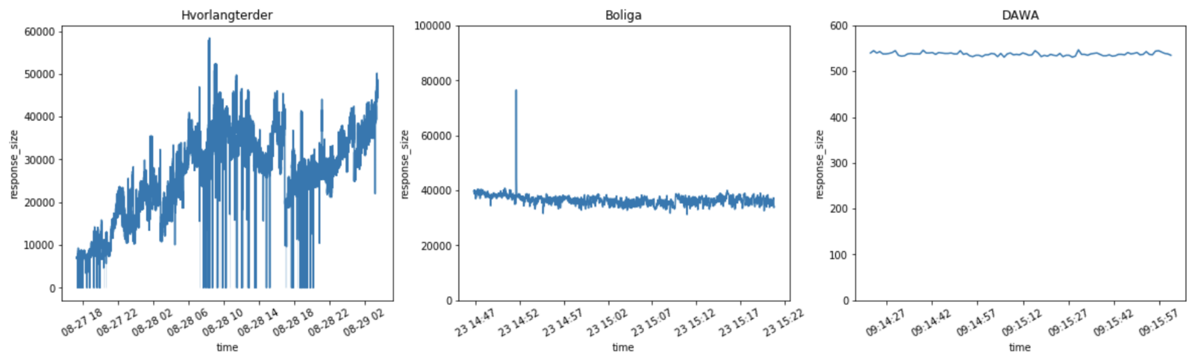
Figure 4



Response Sizes

To validate the response quality, an assessment of the response size was performed. Figure 5 plots response sizes throughout the duration of the web scrapings. Boliga contains a single noticeably large response size within short time, while hvorlangterder response sizes vary throughout the scrape. The DAWA scrape resulted in consistent response sizes, and therefore needed no further investigation.

Figure 5



Examining the distribution of sample sizes in figure 6, the Boliga and DAWA response sizes seem to be close to normally distributed, except for two single cases of large sample sizes in the Boliga log. By manually assessing the returned output of the Boliga scrape, these two requests were found to return a larger amount of data than usual. The requests were made to two pages, where a few properties contained nested dictionary attributes approximately 59 times larger than the average size of dictionaries within the 'image' data column. These dictionaries were image descriptions, and were not utilised within this research. These larger responses were deemed as not polluting the results of the research.

Figure 6

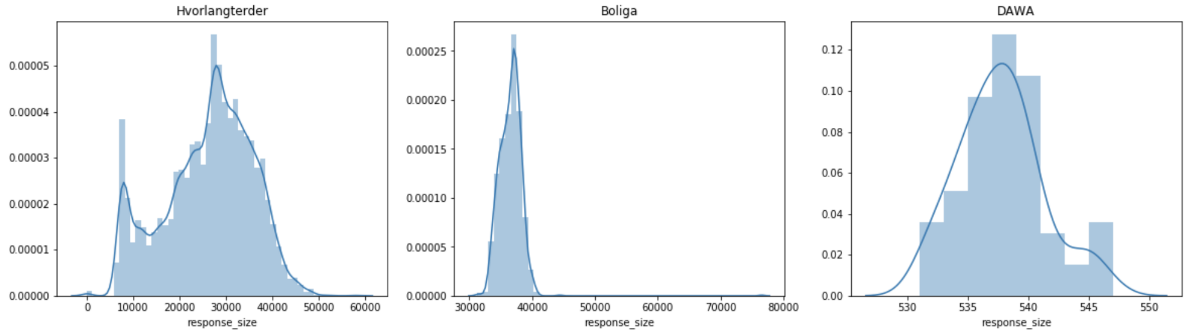


Figure 6 depicts that the hvorlangterder response sizes were a bit left skewed, while Figure 5 shows that the distribution of large sizes happens during the middle of the scrape. With a further glimpse into the log, we observed two things. Firstly, all response sizes of 0 are related to the errors, displayed in figure 2. The variance of responses could be explained by the type of data retrieved from the scrape. The request was set to return names, addresses, distances and other information, which sizes are varying in nature. When extracting the value from every single key of the dictionary output, no errors or missing values were found. Further control of samples of large and small sizes was performed without discovering flaws. The response sizes from hvorlangterder are found to vary in nature, which may explain the difference.

Assessment of the Scraping Logs

The individual analysis of the scraping logs showed minor suspicious traits which were assessed manually through the output or logs. The most noticeable analysis being differences in response sizes and response times from the hvorlangterder scrape. This scrape consumed large amounts of time and computing power, as every single observation from the Boliga data, had to be input for requesting a response. In an extensive research, these problems could be handled by gaining direct access to the website instead of scraping and would benefit from running on a server with greater computing power. The results of the data collection were evaluated to be satisfying, and not corrupting the results of our research, as fairly reasonable explanations for noticeable dissimilarities within the logs were found.

Critique of Logs Using Scraping Class

In the process of analysing the scraping logs, minor flaws from the scraping library named 'scraping_class' were detected. The returned columns 't' and 'delta_t' contained non-generic formats of epoch time. This was fixed by applying string manipulation and datetime conversions.

3.4 Merging Data

Pandas objects can be combined in different ways according to the nature of the features in the dataset. Relational database style operations are based on linking keys together, thereby maintaining the relationship between the combined datasets[21].

The data collection and scraping process provided a total of 15 Datasets from Boliga, hvorlangterder, DAWA, the Danish Police, Social- & Indenrigsministeriet and Danmarks Statistik. This section will describe how each dataset was merged, using pandas relational database styled merge and join operations.

Boliga.dk

An evaluation of the Boliga datasets features and their ability to support further data collection led to the utilisation of the following features:

- [Longitude, Latitude]: Geographical placements of the properties
- Municipality: a numeric code for municipalities in Denmark

The Boliga dataset acted as a master dataset and was joined or merged upon throughout the process, and acted as the *left* of all operations. The Longitude and Latitude features served as specific coordinates for valued properties but were in few cases repetitive regarding different apartments from the same complex. The municipality code was used for translational purposes between the master dataset and other datasets.

DAWA

The DAWA dataset was scraped with an input of a distinct municipality code from the master dataset, returning a dataframe of municipality names for each code. This dataframe was merged onto the Boliga dataframe using a many-to-one merge with municipality codes as the merge keys. The scraped municipality names from DAWA were used as the merge keys for all further merges of municipal data.

Danmarks Statistik, Danish Police and Social & Indenrigsministeriet

Of the municipality-based data collected, the merging of these to our master dataframe was subsequently performed identically. Every dataset contained a column for unique municipality names and values for the given feature of the dataset. A many-to-one left merge was performed with the master dataframe, using municipality names. The result being a master dataframe containing municipal specific features for properties.

The Danish Police datasets contained totals for different categories of crimes reported within municipalities. These datasets were outer joined with an index set to contain municipality names. Afterwards, the values were added to each other, providing a total of reported crimes within each municipality. The police statistics website does not specify whether different types

[21] McKinney, W. (2018). *Data Wrangling: Join, Combine and Reshape*. p. 231

of reported crimes can relate to a single case, but we assessed that the number of crimes reported provides a meaningful feature either way. The dataset containing total reported crimes per municipality was merged with the master dataset in the same fashion as previous municipality-based features.

Hvorlangterder.dk

The hvorlangterder scrape provided location specific features, taking an input of Latitude and Longitude coordinates. The scraping function, which was created for returning location-based features, created a column for the row specific values. Therefore, a merge operation was not necessary, but could alternatively have been managed with a one-to-one inner-merge on id.

3.5 Data Cleaning

The scraping of Boliga left us with 65,950 observations. However, big data is often dirty and requires tidying before it can be used in any meaningful statistical context (Salganik 2018). The data cleaning process consists of removing duplicate values, handling missing data and manipulating strings[22]. The pandas library contains functions to handle these data cleaning methods and was used throughout the process. In the following sections we describe our data cleaning efforts.

3.5.1 Initial Clean-Up

The initial clean-up filtered out rows that contained illogical values. Here we focused our efforts on removing all observations that contained a municipality code of zero. Additionally, we chose to exclude any real estate valued below DKK 100, these listings existed on Boliga's website, but were all auction listings. We excluded these listings on the basis that they were unrealistically priced compared to the property's real market value.

Boliga subdivides its listings into ten real estate types. We chose to exclude the typification "other", as there were only 17 houses listed in the category - too few for training our machine learning model. Furthermore, we also removed any listing without coordinates as these were vital for the scraping process of hvorlangterder.dk.

We dropped all observations with an unreasonably high days-for-sale, as we saw these instances to not represent reasonable pricing or demand. We have thus set an arbitrary limit of 3 years (1,095 days), and omitted any observations that has been on the market for longer than that. This results in the omission of approximately 5% of our dataset. The highest mean "days-for-sale" on a municipal level was roughly 600 days, so to avoid discrimination against the observations from the municipalities with a longer average days-for-sale we set the cut-off somewhat higher than the highest mean. Another option was to set the limit according to each municipality's

[22]McKinney, W. (2018). 'Data Wrangling: Join, Combine and Reshape'. In W. McKinney, *Python for Data Analysis* p. 191

mean, a solution that was a bit more time consuming. In table 1 the number of rows filtered at each step of the cleaning can be perused.

Table 1: Filtered Rows

Feature	***n***
Municipality code	231
Price	11
Coordinates	275
Type	17
Days-for-Sale	3,940
Total[23]	4,332

3.5.2 Apartment Lot Size

Inspecting our data we found that out of 8,028 apartments, 1,176 had a lot size. These listings included the apartment complex' common area, whereas the rest did not. To overcome this discrepancy we chose to set all non-zero values to zero, as not to confuse our model with an inconsistent feature[24].

3.5.3 Final Touches

After merging the housing info with the sociodemographic data and the distances from hvorlangterder.dk, we removed a single duplicate and deleted all columns not relevant for our analysis. After the clean-up of the data we were left with a dataset consisting of 61,618 observations, 37 features and 1 target variable. The features are of both categorical and continuous measures.

3.6 Descriptive Statistics

In this section, we will examine the more apparent patterns of our collected data. This is done to familiarise ourselves with the data, before we commence the analysis.

3.6.1 Key Statistics

Some key statistical characteristics are presented in table 2:

[23] The total is the count of how many rows were dropped. Some rows were missing multiple features

[24] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*

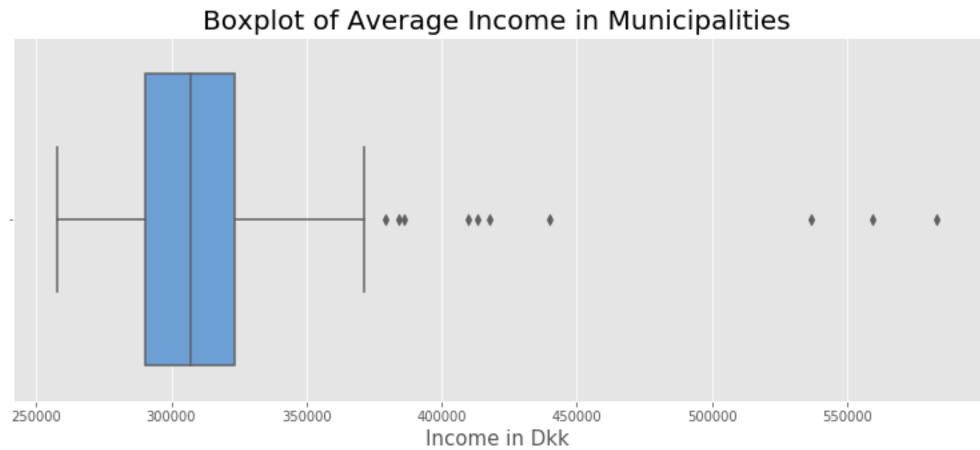
Table 2: Key Statistics

	Price	Rooms	m^2	Average Municipal Income
Mean	2,312,798	4.21	127.82	312,957.03
Std	2,351,127	2.12	75.67	46,095.34
Min	15,000	0	0	257,776
25%	985,000	3	82	290,973
50%	1,695,000	4	123	302,153
75%	2,895,000	5	167	318,745
Max	85,000,000	50	2,390	583,331

Initially, it should be addressed that some of the properties has 0 rooms and consists of 0 m^2. This is due to the fact that we have also included property on which housing has not yet been built. Examining the 25%-quintile and the 75%-quintile of the valuation prices, it becomes apparent that there are some substantial outliers both to the cheaper and expensive side.

It is worth noting, that there is quite a big difference between the lowest average municipal income of DKK 257,776. and the highest of DKK 583,331. This difference becomes further noteworthy when assessing the 75%-quintile. The difference from the 75%-quintile to the highest average municipal income is more than 4 times the difference of the 75%-quintile and the lowest income. The income distribution is visualised in the following figure:

Figure 7

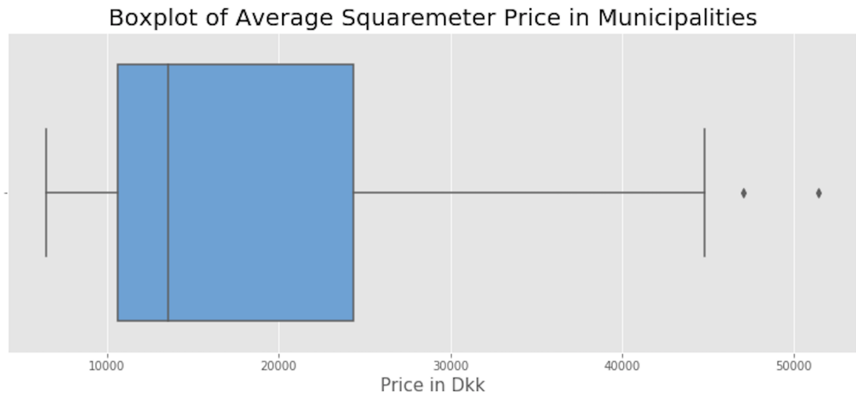


The boxplot illustrates that the distribution is right-skewed, where some municipalities has a sizeably higher average income than the rest of the Danish municipalities.

3.6.2 Prices in Municipalities

We plot the average square meter valuation price in the different municipalities:

Figure 8

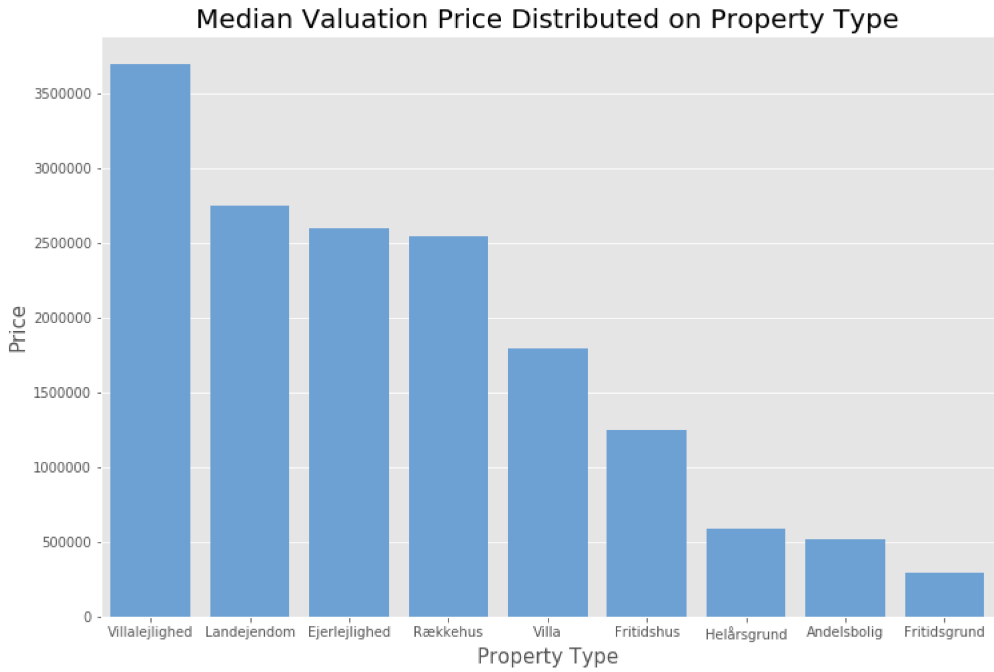


It becomes apparent that the distribution is quite right-skewed and that there are a few municipalities whose price per square meter is much higher than the rest of the municipalities. Recalling the geoplot in figure 1, these expensive outliers are heavily concentrated around- and north of Copenhagen. The distribution looks similar to that of figure 7, though we cannot declare any correlation between average municipal income and average municipal price per square meter. Nonetheless, it becomes apparent that the valuation price of property is highly discriminated by municipal factors. The scope of this research paper is exactly to examine these factors and attempt to use these to valuate an unseen, out-of-sample property.

3.6.3 Property Type

Another worthwhile consideration is that we have included all types of properties. It would be reasonable to assume that there is an average difference in valuation pricing depending on the type of property. Figure 4 displays the median valuation price for each type:

Figure 9



It is interesting to note that the most expensive properties are apartments as opposed to houses. This is not especially surprising though, as it is a well-established trend that real estate prices in major cities are skyrocketing. Refraining from delving deeper into a discussion of global urbanisation, we retain the fact that property type does have an effect on average valuation. We will include the 'type' feature in our impending model training to control for this effect.

4 Methods

The objective of applying Machine Learning (ML) is to train a model that is able to make predictions on new data. By feeding a model labelled data as well as data samples, ML will define the algorithm that perform best predictions[25]. This research paper implements a ML regression prediction model. The following section will establish the methods utilised for optimising the prediction model.

4.1 Fitting the model

The potential problems of underfitting and overfitting should be assessed when fitting a model to data. A model is underfitted if it hardly captures the variation of the sample data. It is then said that the model has *high bias*. A model is overfitted, when it is overly sensitive to the idiosyncrasy of the sample data and captures the variation in too great detail. This problem

[25] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*. p. 3

often comes with the introduction of a sizeable number of features. Overfit models are said to have *high variance*[26]. In both cases, the model will generalise poorly. A key step in defining a decent model in machine learning is to find an optimal bias-variance balance, by tuning the complexity of one's model. This is done through *regularisation*. In this research two different types of regularisation are applied; LASSO and RIDGE.

4.1.1 LASSO

Regularisation by LASSO, will penalise complexity of the model by the sum of the absolute value of the coefficients. This penalty will make the model less complex and more appropriate for prediction[27] [28].
LASSO minimises:

$$L_{LASSO}(\hat{\beta}) = \left(\sum_{i=1}^{n} (y_i - \hat{y}_i(\beta))^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \right) \qquad s.t \quad \lambda \geq 0$$

Another convenient attribute of the LASSO penalty is that some estimates are set equal to zero and thereby produce sparse models[29]. LASSO thereby performs feature selection.

4.1.2 RIDGE

Opposed to LASSO, RIDGE does not force features to be omitted. Instead RIDGE penalises the magnitude of the coefficients. RIDGE minimises:

$$L_{RIDGE}(\hat{\beta}) = \left(\sum_{i=1}^{n} (y_i - \hat{y}_i(\beta))^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^{\,2} \right) \qquad s.t \quad \lambda \geq 0$$

4.1.3 Polynomial Features

Regression models assume linear relationships as standard. One way to account for violations of this assumption, is by utilising polynomial terms within the regression. These polynomial terms create extensive polynomial combinations between the features, fitting the model more precisely to the given data[30]. Polynomial expansion may cause overfitting of the model, generating poor predictions on new data.

[26] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*. p.73

[27] Foster, Ian; Rayid Ghani Ron S. Jarmin, Frauke Kreuter, Julia Lane; *Big Data in Social Sciences, A Practical Guide to Methods and Tools* p. 173

[28] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow* p. 332

[29] Hal R. Varian. *Big data: New tricks for econometrics*. Journal of Economic Perspectives. p.19

[30] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow* p. 334

4.2 Selecting Features

In case regularisation is not sufficient to cope with the overfitting of the model, exclusion of features is a viable approach. With the number of scraped features in this research taking into consideration, a recurring overfitting of the model would be likely. As consequence a deliberate exclusion of features of interest is necessary.

4.3 Optimising the Hyperparameter

To minimise the mean squared errors of our LASSO and RIDGE regression we performed k-fold cross validation to optimise the hyperparameter λ. We split the data into a test set and a development set, consisting of respectively 20% and 80% of the total observations. Subsequently, we use k-fold cross-validation to randomly split the development set into k folds, where k-1 folds are used to train the model. The remaining fold is used to validate the model's generalisability by calculating the mean squared errors of the trained model's prediction of the left-out fold[31]. This process is repeated k times and each time a new fold is left out for validation. Since we are working with a relatively large dataset we chose to split our data into 5 folds, and computed the average MSE for the 5 iterations. By using the k-fold cross-validation method we minimise the concern that the estimation of our model's performance is due to a lucky or unlucky split of the data.
We performed this procedure for different values of λ. Initially we tested greatly different values of λ to prevent convergence issues. On subsequent iterations the value-span was narrowed down. We eventually chose the value of λ which yielded the smallest average MSE over the 5 folds. We both calculated the optimal hyperparameters for a RIDGE regression model and a LASSO regression model.

4.4 Predictive Performance

With the hyperparameters optimised, final model performance can be evaluated. Once more a cross-validation is carried out, which returns the average performance error of each model. By retraining the models on the complete training set and testing on the independent test set, performance measures are obtained[32].
The performances of the models are simply measured by MSE.

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

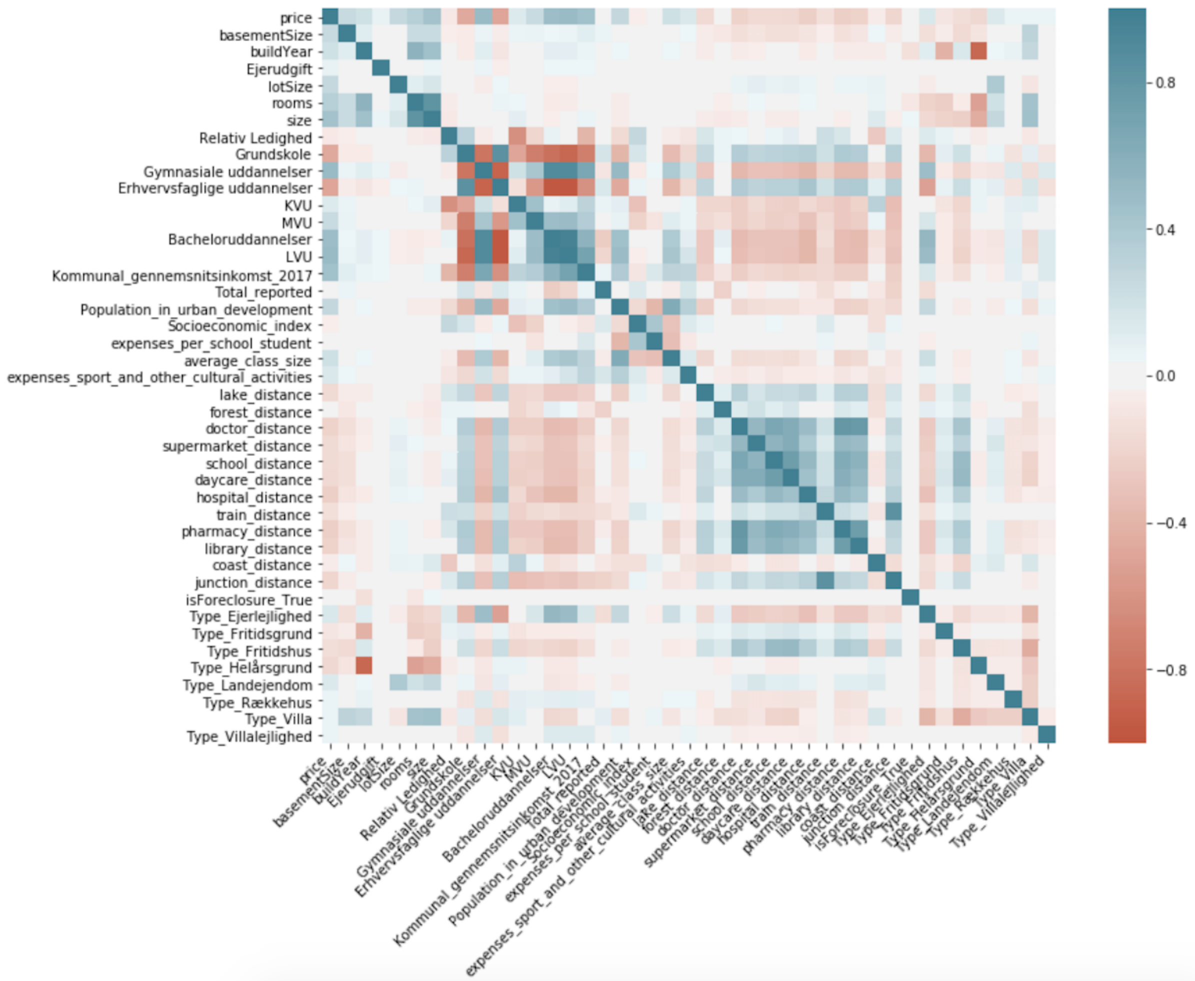
[31] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*. p.191

[32] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*. p.192

5 Analysis

A number of features are excluded before training the model. The exclusion is determined by investigation of the single correlation between the individual features and the target variable, shown in figure 10.

Figure 10



A list of the excluded features is found in appendix 11.0.1. The model used for ML is found in appendix 11.0.2. The initial 5-fold cross-validation to obtain the optimal values of the hyperparameters in each regularisation yields:

Table 3: Optimal hyperparameters, two degrees polynomial features

	λ_{LASSO}	λ_{RIDGE}
2 degrees	372.76	22.76
3 degrees	2310.13	432.88

The prediction-errors of each model, with two and three degrees of polynomial features respectively, are printed below:

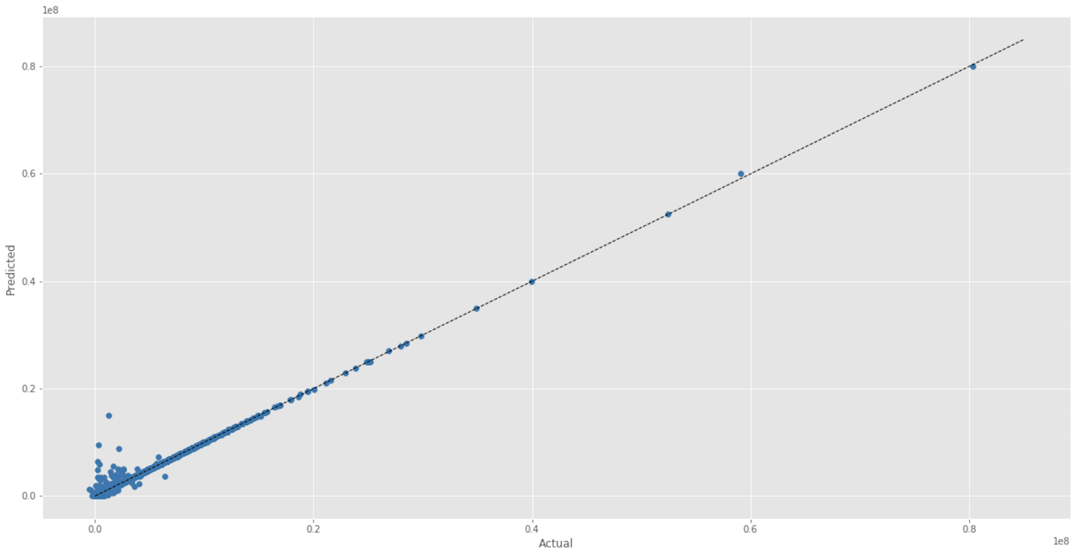
Table 4: Prediction Errors.

	MSE	RMSE	MAE
LASSO 2-deg	59,081,597,331.11	243,067.06	47,660.52
RIDGE 2-deg	62,416,553,466.45	249,833.05	51,306.19
OLS 2-deg	64,297,245,726.80	253,569.02	52,035.52
LASSO 3-deg	60,103,955,188.37	245,161.08	47,126.43
RIDGE 3-deg	55,667,264,851.14	235,939.11	43,630.55
OLS 3-deg	2.848e+28	1.687e+14	2.637e+12

Since our efforts are aimed at predicting the valuation of a out-of-sample real estate property, we want to penalise the size of the error. Hence, the MSE-score is a better indicator for our model's performance as opposed to the mean absolute error (MAE). Since MSE squares the error terms, it gives a relatively high weight to large errors, which is desirable for our current aim.

From table 4 it is evident that the model with 3 degrees of polynomial features regularised by RIDGE regression performs best. Figure 11 shows a plot of the predicted values of our model against the actual values of the test data.

Figure 11

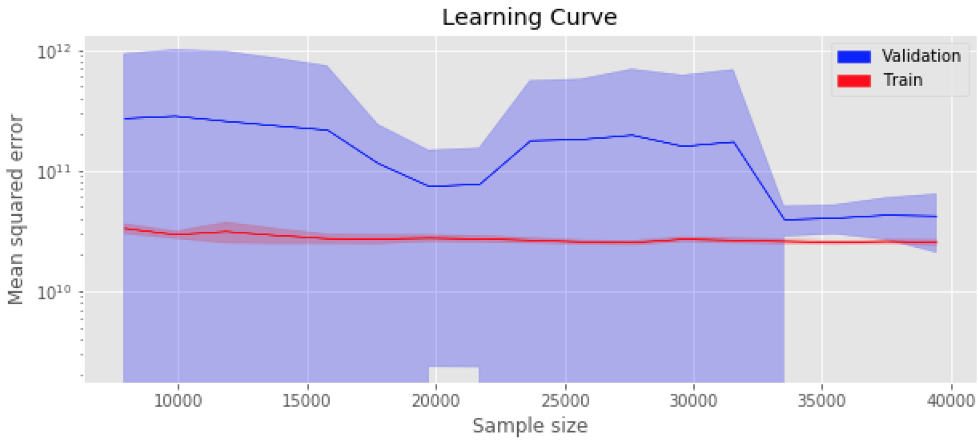


We are somewhat satisfied with the average performance of the model. With an MSE of approximately $55 \cdot 10^{10}$ and a MAE of DKK 43,630.55, we find that the model performs quite well on unseen data. Therefore, we chose this model as the final model from which we will conduct the rest of the analysis with. In the following section further assessment of the model will be conducted

Learning Curve

We examine the learning curve of the model to assert whether the model suffers from immediate over- or underfitting problems and whether these would be remedied by collecting more data[33] The following plot illustrates the learning curve of our model:

Figure 12



The plotted curves represent the average performance of the validation- and training sets, while the width of the plotted curves expresses the 95%-confidence interval of the performance. The performance on the validation set is very inconsistent and varies a lot. Nonetheless, it seems as if the gap is decreasing between the performances of the training- and validation sets as the sample size increases. At around 33,000 observations the validation curve drastically decreases and both curves flatten out. Though the confidence interval of the validation performance is still noteworthy, it also improves dramatically when the sample gets sufficiently high. Just before 40,000 observations, the confidence interval of the validation curve actually drops below the train curve, which should not be possible. We have not found a meaningful explanation aside from noise or chance [34] All in all, the learning curve would indicate that the performance of the model will not benefit from additional training data.

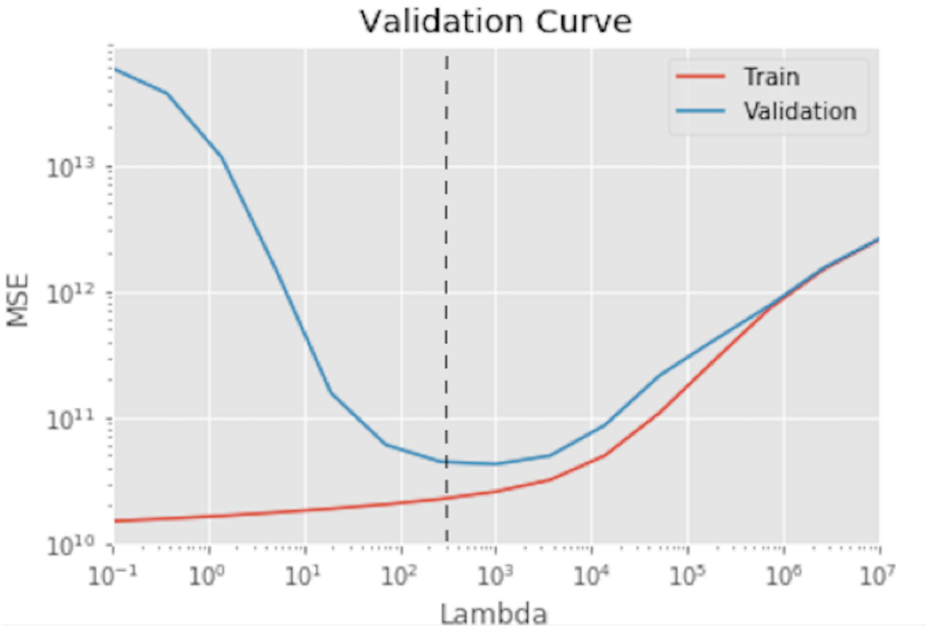
Validation Curve

We examine the validation curve of the model to see whether we have found a good bias-variance trade-off with our optimised hyperparameter. We plot the average train and cross-validation performance for different values of λ:

[33] Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow* p. 196

[34] jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html

Figure 13



The dotted line approximately represents the optimised value of λ from our previously conducted k-fold cross-validation. The curves suggest that we have found the optimal degree of regularisation with our hyperparameter. Had we chosen a smaller value of λ, the model would have had too high variance and thus been overfit. This is deduced by how poorly the validation predictions performs at smaller λ, as opposed to the training performance. On the contrary, had we chosen a greater value of λ, the model would have been too biased. This is illustrated by how poorly the training data, as well as the validation data, performs at higher values of λ . In that case, we would have underfitted our model. Nonetheless, it is apparent that our model performs better on the training data, which indicates that it retains a degree of overfitting. For the abovementioned reasons, it would not be meaningful to minimise this performance gab by adjusting the hyperparameter. Other options for decreasing the variance of the model, could be to include more training data, but as illustrated by the learning curve, this would not seem to effect the variance. Lastly, as overfitting is often caused by a sizeable number of features, omission would be a reasonable mean. As we have already conducted feature selection on the basis of correlations, we did not find a feasible way to conduct further selection without omitting theoretically warranted features. This could be a subject for further studies, to investigate whether this could improve the performance and evaluation of the model.
To sum up it would seem that we have found the hyperparameter which yields the optimal bias-variance trade-off. We can thus conclude that we have optimised our model under the given resources.

6 Discussion

6.1 Alternative Regularisation

As an alternative method of regularisation, we could also have used a regularisation regression called Elastic Net. It is a mix of both LASSO and RIDGE regressions, and minimises[35]:

$$L_{elasticnet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i}(\beta))^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\hat{\beta_j}^2 + \alpha\sum_{j=1}^{p}|\hat{\beta_j}|\right)$$

$$0 \leq \alpha \leq 1 \quad \wedge \quad \lambda \geq 0$$

Note that an Elastic Net regression contains 2 hyperparameters, which also makes it additionally difficult to compute and optimise for big datasets like ours. Optimising the hyperparameters can be done through Grid Search which is a computationally demanding method. We did attempt this method on a subsample of our data, which caused our kernel to crash. With a smaller dataset or with a stronger computer, the performance of an optimised Elastic Net regression should have been taking into account when deciding on a model. As we did not find a lucrative way to optimise the model, we have not included it.

6.2 Best Subset Feature Selection

The current feature selection consisted of selecting features based on their individual correlation to the target. Although providing a satisfying performance, the feature selection could be executed better. The optimal approach to our feature selection, would be applying a Best Subset Selection approach. The Best Subset Selection takes all variable combinations into consideration of the given model and selects the best performing model. Although this approach would provide the best feature combination, the process is highly time consuming [36]. LASSO-models eliminate features irrelevant to the prediction, in contrary best subset feature selection provide a more precise dataset for testing more models.

6.3 Data Critique

An interesting prediction which could have been done using the same methods, would have been to predict selling prices. This could have been done simply by scraping data on sold housing. This could be of more value for private agents, whose main interest would be the actual selling price instead of the valuation.
In a prediction-model like this it is near impossible to evade some form of omitted variable bias.

[35]Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow* p. 332

[36]Gareth, J., Witten, D., Tibshirani, R., & Hastie, T. (2017). Linear Model Selection and Regularisation. In J. Gareth, D. Witten, R. Tibshirani, & T. Hastie, *An Introduction to Statistical Learning*, p. 207

A significant amount of potentially important factors can not be acquired. For example the view from the listed housing will for sure be of great impact of the valuation price. Another factor of interest could have been an evaluation of the condition of the housing. Unfortunately the statement of property[37] are not publicly accessible.
The prediction of the cooperative housing valuation[38], are subject to significant bias, since the cooperative housing that enters the market through a realtor often would be those with critical amount of undesirable characteristics. Future studies could exclude cooperative housing in an effort to increase the prediction capabilities.

6.4 Model Limitations

The exclusion of municipality dummies as regressors in the prediction models, leaves the model to only separate municipalities by the sociodemografic factors, which are subject to low variance. This makes distinction between municipalities a lot more difficult as well as uncertain. Without the municipality dummies, the interpretation of the prediction results are more insecure. Geographical boundaries could have been limited way more, resulting in easier interpretation of predictions. Potentially by predicting only for a specific city or municipality.
If one intended to predict valuations country-wide more realisticly, a model for each municipality could have been another interesting approach, resulting in 98 separate models. The evident downside of this approach being the prospect of limitations in available data since some municipalities has very few listings. Another approach could have been to group cities which share characteristics. Ex. The three biggest cities, the islands, the country-side towns.

6.5 Model Selection

Instead of only assessing the different models by MSE, the information criteria of the respective models could have been introduced, which could have altered the selection of preferred model. Both AIC and BIC would likely have suggested LASSO with 2-degrees of polynomial features as preferred model, as both criteria are increasing with the numbers of estimated parameters[39].

7 Conclusion

In this research paper we have used machine learning to train a model to predict valuation prices of real estate in Denmark. The model train on sociodemographic and geographical data.
We did this by scraping Boliga.dk for all of their listed housing and combining this with municipal data on income, education, schools and urbanisation. Additionally, we also scraped data from hvorlangterder.dk which calculates the distance from an address, to different every-day

[37] red. Tilstandsrapport
[38] http://housingpeople.dk/en/housing-guide/housing-types/cooperative-andelsbolig/
[39] Peckov, Aleksandar. (2012) *A Machine Learning Approach To Polynomial Regression*

commodities such as hospitals, schools and shopping. We have explained how we carried out our gathering of data, how we have structured the data and how we cleaned it.
We performed descriptive statistics to gain a better acquaintance with the data. We conducted feature selection by assessing the correlation between the different features and the target variable, which in our research has been the valuation price. We trained both a LASSO- and a RIDGE regression model and compared their performances with the performance of a standard linear regression. Through k-fold cross-validation, we estimated the optimised hyperparameters for both regularisation models. We found that the RIDGE regression with three polynomial features performed best, at predicting out-of-sample data with the performance measure being the smallest mean square error. When predicting the test data the model had a MSE of 5.566e+10 with a optimised hyperparameter λ=432.88 which we found to be a decent performance. We have assessed the final model and examined to which degree we have managed to find a good bias-variance trade-off for our model and concluded that our model suffers from a degree of overfitting. This was done by examining the validation- and learning curve of our final model. Additionally we addressed possibilities of further optimising the model.
Finally, we have discussed the results of our analysis. We have critically assessed the boundaries of our research and drawn up potential possibilities for improvement.

8 Litterature

Foster, Ian; Rayid Ghani Ron S. Jarmin, Frauke Kreuter, Julia Lane; *Big Data in Social Sciences, A Practical Guide to Methods and Tools*, CRS Press 2017

Rashka, Sebastian; Mirjalli, Vahid; *Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*, 2^{nd}editon, Packt Publishing 2017.

Varian. Hal R. *Big data: New tricks for econometrics*. Journal of Economic Perspectives, 28(2):3–28, 2014.

Consumer Data Research Center, UK [CDRC] (2018) *The General Data Protection Regulation & Social Science Research* [Online]: ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020_eth soc-science-humanities_en.pdf [Accessed on 25/08/2019]

European Commission (2018). *Ethics in Social Sciences and the Humanities* [Online]: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020_ethics-soc-science-humanities_en.pd [Accessed on 25/08/2019]

Popper, Karl (1969) *The Moral Responsibility of the Scientist*. Encounter, March 1969, pp. 52-56 [Online]: http://www.unz.com/print/Encounter-1969mar-00052 [Accessed on 25/08/2019]

Rotblat, Joseph (1999) *A Hippocratic Oath for Scientists*. Science, November 1999: Vol. 286, Issue 5444, pp. 1475 [Online]: https://science.sciencemag.org/content/286/5444/1475.full [Accessed on 25/08/2019] DOI: 10.1126/science.286.5444.1475

Salganik, Matthew J. (2018) *Bit by Bit Bit - Social research in the digital age.* Princeton, NJ: Princeton University Press.

Sample, Ian (2019, Fri 16 Aug 2019) *and tech specialists need Hippocratic Oath, says academic.* The Guardian [Online]: [Accessed on 25/08/2019]

Vox.com (2018) *The Cambridge Analytica Facebook scandal* [Online]: [Accessed on 26/08/2019]

Boligagruppen.dk (2019) *Om boliga gruppen* [Online]: www.boligagruppen.dk: [Accessed on 27/08/2019]

Shiab, Nael (2015) *On the Ethics of Web Scraping and Data Journalism. Global Investigative Journalism Network.* [Online]: gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/: [Accessed on 27/08/2019]

McKinney, W. (2018). 'Data Wrangling: Join, Combine and Reshape'. In W. McKinney, *Python for Data Analysis*. Sebastopol: O'Reilly Media Inc.

Athey, Susan (2018) 'The Impact of Machine Learning on Economics' in Ajay Agrawal et al. (eds) *The Economics of Artificial Intelligence: An Agenda* (2019). Chicago: University of Chicago Press.

Einav, Liran and Jonathan Levin (2014) *Economics in the age of big data*. Science, vol. 346, November.

Grimmer, Justin (2015) *We are all social scientists now: how big data, machine learning, and causal inference work together*. PS: Political Science & Politics, vol. 48, January: 80-83.

Lazer, David and Jason Radford (2017) *Data ex Machina. Introduction to Big Data. Annual Review of Sociology*, vol. 43, August.

Mullainathan, Sendhil, and Jann Spiess (2017) *Machine Learning: An Applied Econometric Approach*. Journal of Economic Perspectives, vol. 31 (2): 87-106.

Anderson, Chris (2008) *The end of theory: The data deluge makes the scientific method obsolete*. Wired, 16-07 [Online] www.wired.com/2008/06/pb-theory/ [Accessed on 27/08/2019]

Peckov, Aleksandar. (2012) *A Machine Learning Approach To Polynomial Regression*. Available at kt.ijs.si/theses/phd_aleksandar_peckov.pdf. Doctoral Dissertation. Jozef Stefan International Postgraduate School. Ljubljana, Slovenia, October 2012. [Accessed on 29/08/2019]

(https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html). [Accessed on 29/08/2019]

Gareth, J., Witten, D., Tibshirani, R., & Hastie, T. (2017). Linear Model Selection and Regularization. In J. Gareth, D. Witten, R. Tibshirani, & T. Hastie, *An Introduction to Statistical Learning* (pp. 205-262). New York City: Springer.

https://www.dst.dk/da/Statistik/nyt/NytHtml?cid=27979 [Accessed 27/08/2019]

https://www.dst.dk/da/Statistik/nyt/NytHtml?cid=28741 [Accessed 27/08/2019]

9 Appendix

9.0.1

Excluded Features:
municipality, lotSize, unemployment, Total_reported_crime, Socioeconomic_index, expenses_per_school_stu
expenses_sport_and_other_cultural_activities, forest_distance, coast_distance, isForeclosure, owners_expense

9.0.2

The model used for prediction:

$$\hat{Y_i} = \hat{\beta_0} + \sum_{i=1}^{p} \hat{\beta_i}\mathbf{T}_i + \hat{\epsilon_i}, \quad T_i = \prod_{j=i}^{p} X_j^{d_i,j}$$ [40]

p being total number of features, d_i, j denoting degree of polynomial features.

$$\mathbf{X}_j = \begin{bmatrix} basementSize \\ buildYear \\ rooms \\ size \\ primary\_school\_educ \\ high\_school\_educ \\ vocational\_educ \\ SHE \\ MHE \\ bachelors\_degree \\ LHE \\ avg\_municipal\_income\_2017 \\ Population\_in\_urban\_development \\ Socioeconomic\_index \\ lake\_distance \\ doctor\_distance \\ supermarket\_distance \\ school\_distance \\ daycare\_distance \\ hospital\_distance \\ train\_distance \\ pharmacy\_distance \\ library\_distance \end{bmatrix}$$

[40] Peckov, Aleksandar. *A Machine Learning Approach To Polynomial Regression* p. 6-9

9.0.3 t