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

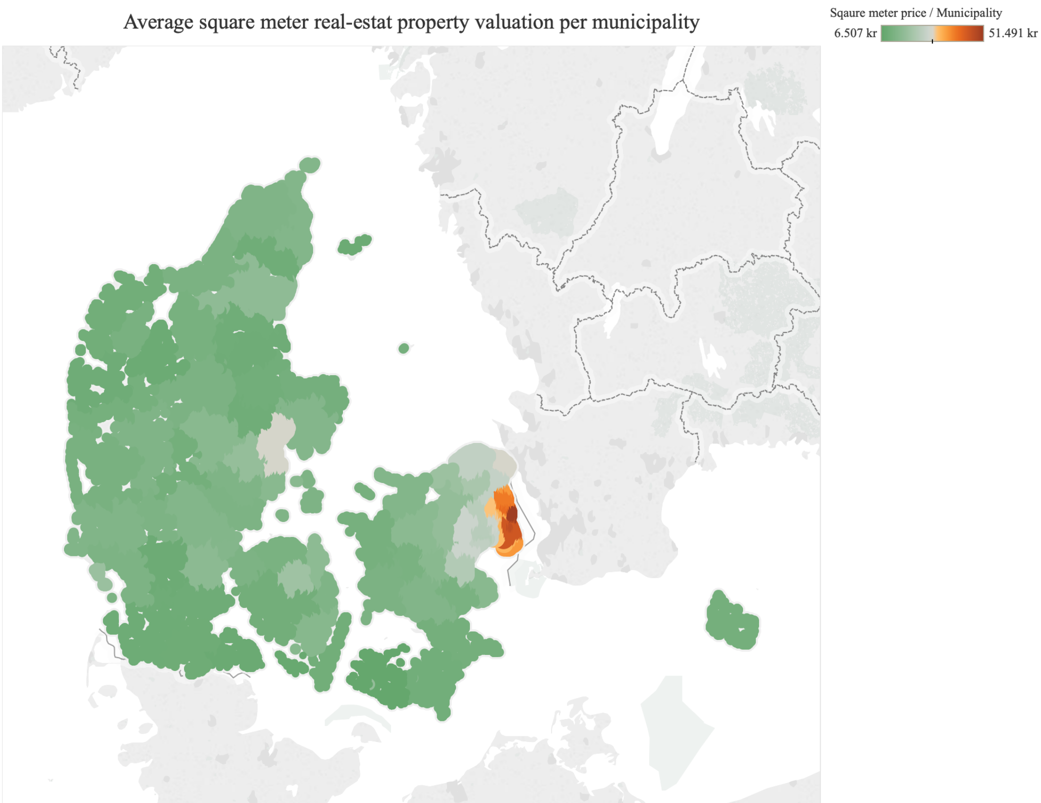
Typeenheder:

Contents

1 Introduction

This report depictss a research performed regarding active Danish real-estate property valuations from 2016-2019. Since 2011, the yearly 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 agents, a closer analysis of current and active valuations add 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 lowest average square meter price valuations. A tendency visualized in the figure, is that average square meter price valuations are higher in 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 15,000 to 85 million. DKK. This research paper intends to analyze real estate price valuations and which effects geo- and sociodemographic criteria affect valuations, using machine learning models. Our research questions is as follows:

Which features are most relevant for predicting evaluations of real estate properties 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 machine learning model. Also, a section regarding the choice and optimization of machine learning models is included, where the intention is to provide insights on our progress on finding the optimal model. As a result, a preferable machine learning model is chosen with a discussion of its usability.

2 Literature Review

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

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 of quantities of data[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 as 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 to understand that big data is not a naturally occurring system, but driven by the engineered purpose of the system. This algorithmic confounding forces 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 various 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].

The technical and theoretical challenges faced by big data and machine learning research are

[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*

important to consider, when employing the tools they provide. One of the major discussions revolve around machine learning's predictive capabilities. Chris Anderson[4] (2008) argues that since the computing power and the scale of data has increased exponentially, our reliance of 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 social sciences greatest contribution to big data research, comes from the organized 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 utilize 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[8].

3 Data Description & Ethics

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 Comissions[9] ethical guidelines and principles for ethical conduct in social data science, the potential harm to users of Boliga's web-page were carefully considered. As a step to prevent the mosaic effect and in an effort to anonymize the scraped data only aggregated data will be presented in this paper, so that no single observation can be identified from the analysed data.

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

No informed consent has 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, sometimes, 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." [10] 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. 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 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, which causes each scraping iteration to pause for 0.5 seconds before commencing on scraping the next page (see appendix XX). Furthermore, Boliga prohibits the use of automated scrapers and bots, which we did not do, as our scraping was done in a specified time-frame and we did not automate the procedure to be done on multiple occasions.

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 labelled XX and XX.

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 "liggetid", price-development and access to BBR - the Danish Building and Housing Register [11]. Giving us unique insights into the pricing of real-estate in all 98 municipalities of Denmark. At the time of scraping 65,950 properties were for sale.

The scraping process was conducted on Friday the 23rd of August 2019. We had advised Boliga of our intent to collect data from their website via our scraper in an effort to identify ourselves and our intent[12]. In order to scrape the data of interest we familiarized ourselves with the

[10] Salganik, Matthew J. (2018) *Bit by Bit Bit - Social research in the digital age.* p.303

[11] www.boligagruppen.dk

[12] Shiab, Nael (2015) *On the Ethics of Web Scraping and Data Journalism.* Global Investigative Journalism Network.

HTML-structure of Boliga. On the basis of these insights we constructed a code, which were able to scrape every page, containing information pertaining 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 URL, 34 features and the target variable (price) was collected. Table 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.

Features Obtained:

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

Hvorlangterder.dk

hvorlangterder.dk returns distances from a given address, to conveniences 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. It constructed a URL (65,950 URL request), which was scraped and the json response was returned as a dictionary of distances to the points of interest. The values from each key in the dictionary was then extracted as a new column in a Pandas DataFrame.

As Jupyter performs poorly at running long asynchronous tasks and estimated a running time of 18 hours this procedure was run in Visual Studio.

Features obtained:

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

Social and economic factors

Social and economic factors on municipality level is collected from statistik.politi.dk and Danmarks Statistik respectively. These factors include income, reported crime, level of highest completed education etc. These are 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 Log Analysis

3.4 Merging Data

Pandas objects can be combined in different ways according to the nature of features in the dataset. Relational database style operations are based on linking keys together, thereby maintaining the relationship between the combined datasets[13]

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 utilization of the following features:

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

The Boliga dataset acts as a master dataset and was joined or merged upon throughout the process, and acts as the "left" of all operations. The Longitude and Latitude features serve as specific coordinates for valued properties but are in few cases repetitive regarding different apartments from the same complex. The municipality code was used for translation 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 as a many-to-one merge with municipality code as the merge keys. The scraped municipality names from DAWA was used as the merge key for all further merges of municipal data.

Danmarks Statistik, Danish Police & Social og 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.

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

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 webpage does not specify whether different types 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 id. As few coordinates regarding apartment properties are repetitive, these are not suitable for merging upon.

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). In the following sections we describe our data cleaning efforts.

3.5.1 Initial clean-up

The initial clean-up sorted out rows that contained an illogical value. Here we focused our efforts on the removing any and all observations that contained a municipality code of 0. Additionally, we chose to exclude any real estate valued below 100 DKK, while these listings existed on Boliga's website, they were all listings, where the seller wanted to sell to the highest bidder. 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 on hvorlangerder.dk.
We dropped all observations with an unreasonably high "liggetid", 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 "liggetid" on a municipal level was roughly 600 days, so as not to discriminate against the observations

from the municipalities with a longer average 'sales time' we set the cut-off somewhat higher than the highest mean. Another option was to set the limit according to each municipality's mean, a solution that was a bit more time consuming, and would have yielded approximately the same result, hence we opted not to do it. In figure XX the number of rows dropped at each step of the cleaning can be perused.

Feature	*n*
Municipality code	231
Price	11
Coordinates	275
Type	17
Liggetid	3,940
Total[14]	4,332

3.5.2 Ejerlejlighed's lot size

Inspecting our data we found that from 8,028 apartments 1,176 of these 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 (Raschka & Mirjalili 2017).

3.5.3 Final touch-ups

After merging the housing info with the demographic 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 and 37 features and 1 target variable.

3.6 Descriptive Statistics

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

3.6.1 Key Statistics

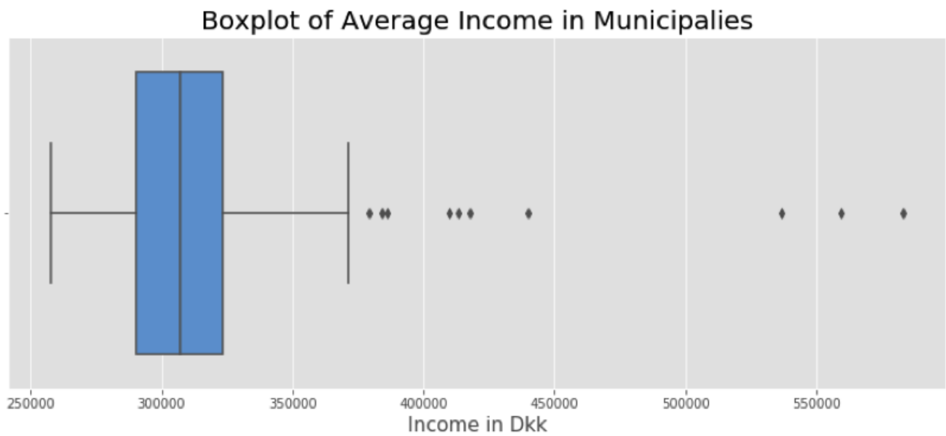
Some key statistical characteristic is presented in table 1:

Table 1: Key Statistics

	Price	Rooms	m^2	Average Municipal Income
Mean	2,312,798.00	4.21	127.82	312,957.03
Std	2,351,127.00	2.12	75.67	46,095.34
Min	15,000.00	0	0	257,776
25%	985,000.00	3	82	290,973
50%	1,695,000.00	4	123	302,153
75%	2,895,000.00	5	167	318,745
Max	85,000,000.00	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 land on which housing has not yet been build. 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 to expensive side. This is also the case with rooms and m^2, where the highest values are sizably higher than the 75%-quintile. 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 2

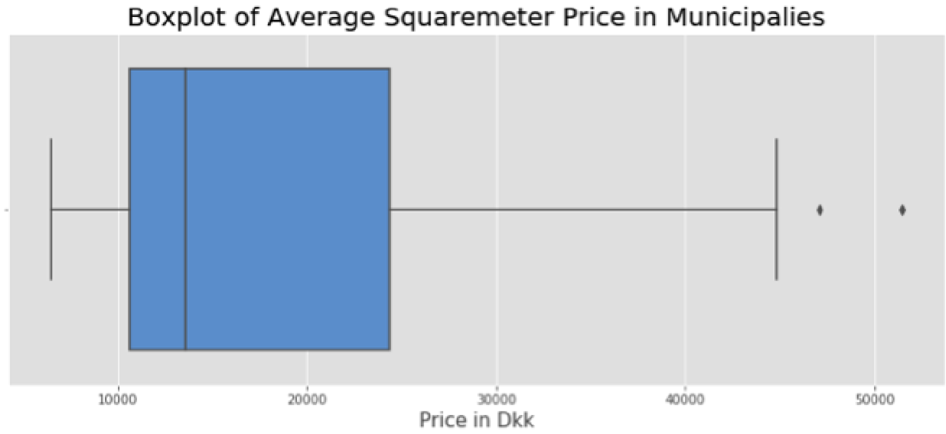


The boxplot illustrates that the distribution is heavily left-skewed, where some municipalities has a sizably higher average income than the rest of the Danish municipalities.

3.6.2 Prices in Municipalities

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

Figure 3

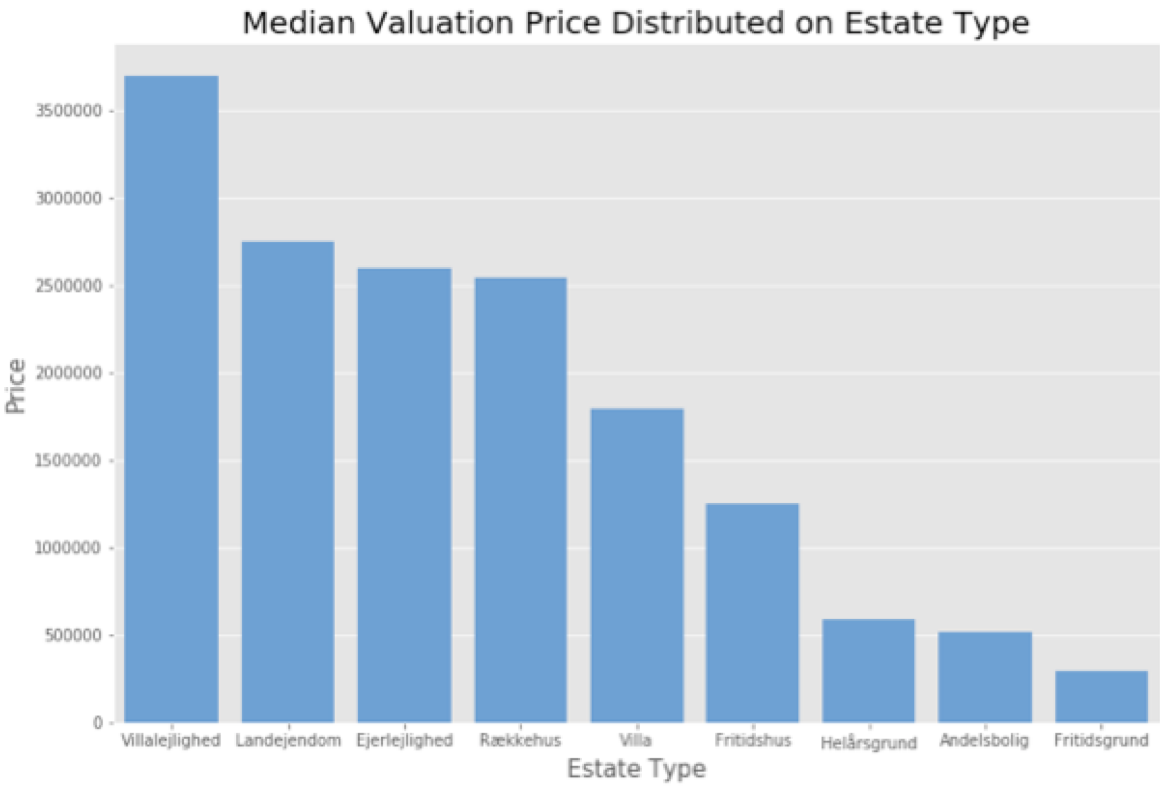


It becomes apparent that the distribution is quite left-skewed and that there are a few municipalities who's price per square meter is much higher than the rest of the municipalities. Recalling the geoplot in figure 1, these expensive outliers are heavily concentrated around- and north of Copenhagen. The distribution looks similar to that of figure 2 (The previous boxplot), 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 assignment is exactly to examine these factors and attempt to use these to evaluate 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 4



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 to a discussion of global urbanization, we retain the fact that property type does have an apparently significant effect on valuation pricings on average. We will include the 'type' feature in our impending model training to control for this effect.

3.6.4 The Dataset

The merged dataset contains 61.618 observations and 37 features. The features are of both categorical and continuous measures. Skal dette ikke bare op i Merging data?

4 Methods

4.1 Supervised Machine Learning

The objective by applying Machine Learning is to train a model that are able to make predictions in near future or never seen data. By feeding a model with labeled data as well as data samples, ML will define the algorithms that predicts the best[15]. This project, implements a ML regression prediction model to predict pricing of housing listed in near future.

4.2 Fitting the model

The potential problems of underfitting and overfitting should be assessed when fitting a model. 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 often comes with the introduction of a sizeable number of features. Overfit models are said to have *high variance*[16]. In both cases, the model will generalize 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 *regularization*. In this project three different types of regularization are applied; Lasso, Ridge and Elastic net.

4.2.1 Lasso

Regularization by Lasso, will penalize 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. [17] [18]
Lasso minimizes:

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

Another convenient attribute of the Lasso penalty is that some estimates are set equal to zero and thereby produce sparse models[19]. Lasso thereby performs the feature selection automatic.

4.2.2 Ridge

The ridge model penalizes with the sum of squared coefficients. Opposed to Lasso, Ridge do not force features to be omitted. Instead Ridge penalizes the magnitude of the coefficients.

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

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

[17] 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

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

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

Ridge minimizes:

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

4.2.3 Elastic-net

The Elastic net combines the penalizing terms of Lasso and Ridge, with α defining the relative weights between Lasso and Ridge. Elastic net minimizes the function:

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

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

4.3 Selecting Features

In case regularization is not sufficient to cope with the overfitting of the model. Exclusion of features are a viable approach. With the number of scraped features in this project, taking into consideration. A recurring overfitting of the model would be likely. As a consequence a deliberate exclusion of features of interest will need to be carried out.

4.4 Optimizing the Hyperparameter

To minimize the mean squared errors of our Lasso and Ridge regression we performed k-fold cross validation to optimize 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 generalizability by calculating the mean squared errors of the trained model's prediction of the left-out fold[20]. 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 relieve ourselves of the concern that the estimation of our model's performance is simply due to a lucky or unlucky split of the data.
We performed this procedure for 12 different values of λ spanning between $10^(-4) and 10^4$. We chose the value of λ which yields the smallest average MSE over the 5 folds. We both calculated the optimal hyperparameters for a Ridge regression model, a Lasso regression model and

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

an Elastic Net regression model. The following table shows the performance of the different models, when trained with their optimal hyperparameter and predicting the test data.

4.5 Predictive Performance

With the hyperparameters optimized, final model performance can be evaluated. Once more a cross-validation is carried out, which return 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[21].
The performances of the models are simply measured by MSE, RMSE and MAE.

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

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

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

5 Analysis

A number of features are excluded before training the model. The exclusion was determined, quite simply, by investigation of the correlation between the complete set of features and the target variable. A list of the excluded features are found in appendix 11.0.1. The model used for ML are found in appendix 11.0.2.

The initial 5-fold cross-validation to obtain the optimal values of the hyperparameters in each regularization yields.

Table 2: Optimal hyperparameters, two degrees polynomial features

	λ_{Lasso}	λ_{Ridge}	$\lambda_{ElasticNet}$
2 degrees	372.76	26.83	1
3 degrees	2310.13	432.88	1

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

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

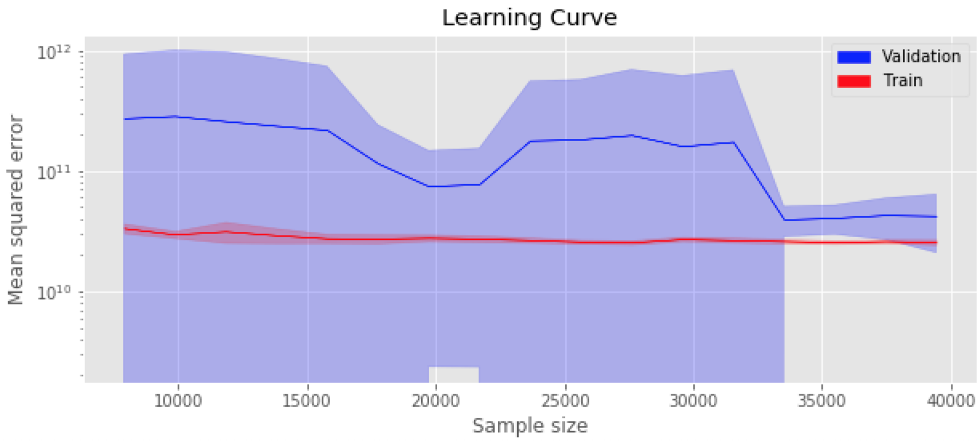
Table 3: Prediction Errors.

	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
Elastic net 2-deg	1	1	1
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
Elastic Net 3-deg	1	1	1
OLS 3-deg	2.848e+28	168,763,945,674,352.3	2,637,476,978,906.39

The Ridge model with three degrees of polynomial features are selected as the best performing model. In regard to MAE as well as MSE.
We examine the validation curve, figure XX, of the model to examine whether we have found a good bias-variance-balance with our optimized hyperparameter. We plot the average train and cross-validation performance for different values of λ:

Learning Curve We examine the learning curve of the model to examine whether the model suffers from immediate over- or underfitting problems and whether these would be remedied by collecting more data[22] The following plot illustrates the learning curve of our model:

Figure 5



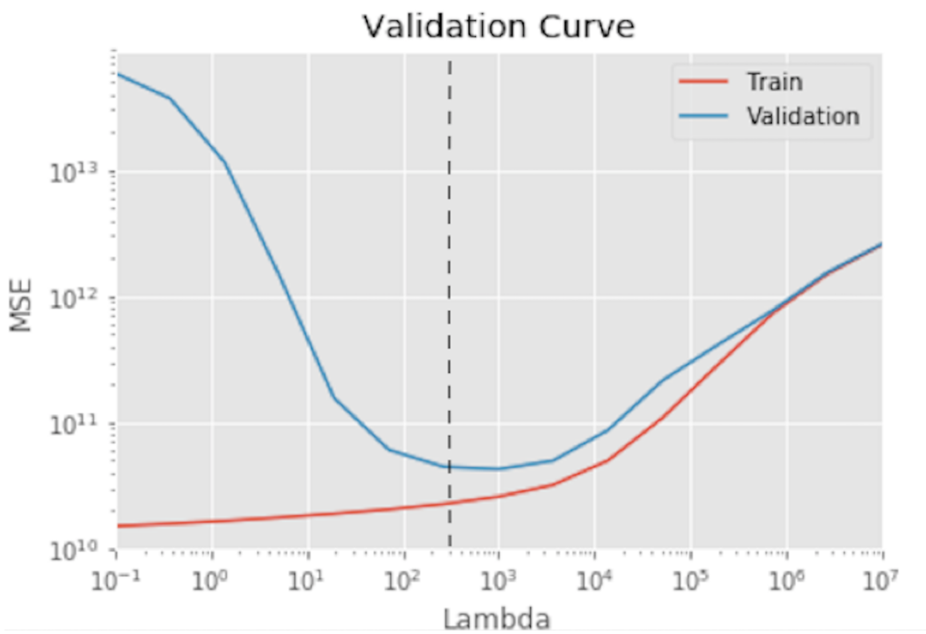
The plotted curved represent the average performance of the sets, while the width of the plotted curves expresses the 95%-confidence interval of the performance. Interpreting the graph could appear quite daunting, as the performances on the validation set are very inconsistent and varies a lot. Though inconsistent, it seems as if the gap is decreasing between the performances

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

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 [23] 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 optimized hyperparameter. We plot the average train and cross-validation performance for different values of λ:

Figure 6



The dotted line approximately represents the optimized value of lambda from our previously conducted k-fold cross-validation.

Firstly, 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. The curve suggest that we have found the optimal degree of regularization 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 lambdas, 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 underfit our model. Nonetheless, it is apparent that our model

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

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 minimize 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 make for more less 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. This could be a subject for further studies, to investigate whether this could improve the performance and evaluation of the model.

In conclusion, it would seem that we have found the hyperparameter which yields the optimal bias-variance-trade-off. We can thus conclude that we have optimized our model under the given circumstances.

6 Results

Figure 7: Predicted Valuation vs. Actual Valuation

7 Discussion

7.1 Data critique

Another 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 instead. This could be of more value for private agents, whose main interest should be the selling price of their housing.

In a prediction-model like this it is near impossible to evade some form of omitted variable bias. A significant amount of potential important factors can not be acquired. For an 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 a evaluation of the condition of the housing, unfortunately the statement of property[24] are not publicly accessible.

The prediction of the cooperative housing valuation are subject to significant bias, since the cooperative housing that enters the marked through a realtor often would be those with critical amount of undesirable characteristics.

[24]red. Tilstandsrapport

7.2 Model Limitations

The necessity to exclude municipality dummies as regressors in the prediction models. Leaves the model to only separate municipalities by the socio-economic factors, which are subject to low variance. This makes distinction between municipality 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 realistic, a model for each municipality could have been more satisfactory. Resulting in 98 separate models. The evident downside of this approach being the prospect of limitations in available data (some municipalities has very few listings). Another approach could have been to group cities which share characteristics. Ex. The three biggest cities, the provinces, the islands, the country-side towns.

7.3 Model Selection

Instead of only assessing the different models by MAE and RMSE. 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 measurement are increasing in numbers of estimated parameters[25].

8 Conclusion

In this assignment, we have used machine learning to train a model to predict valuation prices of real estate in Denmark. We did this by scraping Boliga.dk for all of their active houses for sale and combining this with municipal data on income, education, schools and urbanization. Additionally, we also scraped data from hvorlangterder.dk which calculates the distance from an address, to different every-day 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 on the data 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 assignment 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. Though k-fold cross-validation, we estimated the optimized hyperparameters for both regularization models. We found that the Ridge regression performed best on average, with the performance measure being the smallest mean square error of the

[25] Peckov, Aleksandar. (2012) *A Machine Learning Approach To Polynomial Regression*

model. We have assessed to which degree we have managed to find a good bias-variance-balance for our model and concluded that our model suffers from a degree of overfitting. This was done by examining the validation curve of our final 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.

9 An Ethical Overview

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[26] [27]. 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 under development in laboratories and tech firms[28] [29]. This discussion is nothing new however, as a tentative reformulation of the Hippocratic oath was introduced by Karl Popper[30], wherein he stressed the importance of professional responsibility, a critical mind, and an overriding loyalty towards the betterment of mankind.
Matthew Salganik offer 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 maximize the possible benefits and minimizing 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 stress 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 17th of March 2018. Where Steve Bannon could reveal that between 2013 and 2015 Cambridge Analytica had exploited a loophole in Facebook's API which allowed the company to harvest profile data from 87 million Facebook

[26]Salganik, Matthew J. (2018) *Bit by Bit Bit - Social research in the digital age.*
[27]European Commission (2018). *Ethics in Social Sciences and the Humanities*
[28]Rotblat, Joseph (1999) *A Hippocratic Oath for Scientists*
[29]Sample, Ian (2019, Fri 16 Aug 2019) *and tech specialists need Hippocratic oath, says academic.*
[30]Popper, Karl (1969) *The Moral Responsibility of the Scientist*

users, without the user's permission and use the harvested data to construct a massive targeted marketing database based on the user's likes and interests[31]. 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 and how these networks and their actors' behavior co-evolve, and the emotional contagion experiment from 2012, where approximately 700,000 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 formalize the procedures involved in the data processing and storing of sensitive and private information (CDRC[32]). These principles include and expand upon the principles found in the Belmont and Menlo Report. The most important consideration, however, must be that even a dataset comprising tens of thousands of observations involve human beings who must be protected from adverse side-effects of the social research. There is considerable evidence that point to the fact that even in anonymized 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).

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

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

10 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]

11 Appendixs

11.0.1

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

11.0.2

The model used for prediction:

$$\hat{Y_i} = \hat{\beta_0} + \sum_{i=1}^{p} \hat{\beta_i}\mathbf{T}_i + \hat{\epsilon_i}, \quad T_i = \prod_{j=i}^{p} X_j^{d_i,j} \quad 33$$

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

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

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