KØBENHAVNS UNIVERSITET
ØKONOMISK INSTITUT

Predicting Valuation Prices of Danish Real Estate Property

Undertitel på opgaven

Group 40.

Exam numbers: 115, 144, 146, 204.

Afleveret den: 30/08/2019

Typeenheder:

Indhold

1 Introduction

2 Data description & Ethics

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[1] 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.

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."[2] 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.

2.1 Data sources

Scraping from **boliga.dk** provides key characteristic such as price, squaremeters, location etc. The API hvorlangterder.dk returns measurements of distances from an address to conveniences

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

[2] Salganik, Matthew J. (2019) *Bit by Bit Bit - Social research in the digital age.* p.281

such as hospitals, schools and shopping. Obtaining these distances for all of the listed housing and merging the two datasets, gives additional arguments to geographical-control. The data from hvorlangterder.dk can allow more micro-oriented differences within a given municipality.
Which otherwise would not have been captured.
Social and economic factors on municipality level is collected from statistik.politi.dk
Danmarks Statistik respectively. These factors includes reported crime, income and level of completed education.

2.2 Data construction

Traffic offences are not considered as a "serious"offence and are therefore excluded in the crime statistics. Socio-economic data is converted to ratios.

2.3 Exclusion of outliers

A few of the listed housing, have been for sale for over a decade. These are considered outliers, which are overly priced considering their characteristics. The outliers are excluded, because they would otherwise only contribute to a bias of overestimation. The total number of excluded housing are approximately 4000.

Danmark					
	Sjælland		Ikke Sjælland		
Type/Region	Hovedstaden	Sjælland	Syddanmark	Midtjylland	Nordjylland
Villa	513,36	459,67	503,62	461,96	
Rækkehus	470,67	404,95	460,93	407,24	
Ejerlejlighhed	482,30	463,94	455,87	449,54	
Fritidsbolig	513,36	459,67	503,62	461,96	
Kolonihave	470,67	404,95	460,93	407,24	
Andelsbolig	482,30	463,94	455,87	449,54	
Landejendom	513,36	459,67	503,62	461,96	
Helårsgrund	470,67	404,95	460,93	407,24	
Fritidsgrund	482,30	463,94	455,87	449,54	
Villalejlighed	513,36	459,67	503,62	461,96	
Lystejendom	470,67	404,95	460,93	407,24	

3 Methods

3.1 Supervised Machine Learning

The objective by applying Machine Learning is to train a model that can predict pricing of housing listed in near future. Two types of predictive models will be used for the sake of returning different outputs: Regression-model to return the pricing as a continuous output. As well as a classification-model for predicting the interval which the price is included in.

3.2 Regularization

The use of a sizeable amount of features leads to a significant risk of over-fitting the prediction model. For the purpose of dealing with over-fitting, regularization needs to be carried out. Regularization can be done with different approaches, in this project three types are applied; Lasso, Ridge and Elastic-net.

3.2.1 Lasso

Regularization by Lasso, will penalize complexity of the model by adjustment of parameter estimates. This penalty will make the model less complex and more appropriate for prediction. [3] [4]

Lasso minimize:

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

Another convenient attribute of the Lasso penalty is that some estimates are set equal to zero and thereby produce sparse models[5]. Lasso thereby performs the feature selection automatic.

3.2.2 Ridge

The ridge model add squared terms of the coefficient as penalty

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

Like Lasso, the Ridge model add bias in exchange for a decrease in variance. Resulting in a better prediction model.

[3] 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

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

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

3.2.3 Elastic-net

The Elastic net is a combination of Lasso and Ridge. Elastic net minimizes the function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{P} \hat{\beta_j}^2 + \alpha \sum_{j=1}^{P} |\hat{\beta}_j| \right) Note : AnderledesiRashka!$$

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

Where α denotes the relative mixing between Lasso and Ridge.

3.3 K-fold Cross-Validation

To get the optimal tuning parameter: λ a k-fold cross-validation is performed. The 10-fold CV, divides the data into 10 subsets

4 Analysis

5 Results

6 Discussion

6.1 Data critique

Another interesting prediction which could have been obtained using the same methods, would have been to estimate expected 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 have 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"red. (tilstandrapport) 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.

7 Conclusion

8 Appendix

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[6] [7]. 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[8] [9]. This discussion is nothing new however, as a tentative reformulation of the Hippocratic oath was introduced by Karl Popper[10], 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 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[11]. Other examples of misuse of data acquired from Facebook include the Harvard-run experiment, where students' data was used to

[6] Salganik, Matthew J. (2019) *Bit by Bit Bit - Social research in the digital age.*

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

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

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

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

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

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. 2019).
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[12]). 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).

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

9 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.

Hal R. Varian. *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/r
soc-science-humanities_en.pdf [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. (2019) *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]