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

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 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 / Lasso og derved variable selection

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, a form of regularization needs to be carried out. The Lasso-feature posses the attributes to do so. 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. [1] [2]
Lasso solves the optimization problem:

$$\hat{\beta}(\lambda) = arg\min_{\beta}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i(\beta)^2) + \lambda\sum_{j=1}^{p}|\beta_j|\right) \qquad s.t \quad \lambda \geq 0$$

[1] Big Data in Social Sciences, Forster et al. p. 173

[2] Python Machine Learning, Sebastian Raschka et al. p. 332

Another convenient attribute of the Lasso penalty is that some estimates are set equal to zero [3]. Lasso thereby performs the feature selection automatic.

3.3 Lars

3.4 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

[3] Hal R. Varian. Big data: New tricks for econometrics. p.19

8 Litterature

Big Data in Social Sciences, Forster et al. 2017

Python Machine Learning, Sebastian Raschka et al. 2017, 2^{nd}editon

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