Predicting Housing Prices in Denmark
A Machine Learning Approach*

Group 23: He SHI[†], Mingjie XING[‡], Qi ZHANG[§]
Department of Economics, University of Copenhagen

August 19, 2023

1 Introduction

This project is intended to build a predictor of residential housing prices in cities in Denmark with a machine learning hedonic model based on interior features such as house area and location, and exterior features such as income level of the neighbourhood, weather condition and cities' economic development level. The project mainly covers major cities in Denmark, e.g., Copenhagen and Aarhus. The frequency of the data will be monthly. The trained model based on 2023 data will be tested on 2022 and 2021 data.

2 Literature Review

There are several papers recently on top journals in real estate economics predicting housing prices with machine learning models:
Deppner et al. (2023) applies non-parametric machine learning model with k-fold cross-validation to examine the U.S. commercial real estate appraisal to shrink the deviations between market values and subsequent transaction prices.
Lin et al. (2023) augments the traditional hedonic model with Gradient Boosting Machine algorithm to predict housing prices in Beijing.

*This project contains XXX words.
[†]Data collection, modelling
[‡]Report drafting, modelling
[§]Data collection, modelling

Tchuente and Nyawa (2022) uses artificial neural networks, random forest, adaptive bBoosting, gradient boosting and K-nearest neighbours algorithms to estimate real estate prices in French cities.

2.1 Contribution

The main contribution of this project includes: 1) creating a comprehensive dataset on Denmark residential real estate market; 2) building a predictor of Denmark housing prices in different cities.

3 Data and Methodology

3.1 Data

We are going to scrape the individual level house prices from boligsiden, a major real estate broker in Denmark covering housing prices as monthly data of 36 months scale and detailed features of houses in different regions with an open data access. We will scrape the data from `https://www.boligsiden.dk/`. Structural features in data includes area, location, age and owner income. Web-scraping of the data is allowed for academic use. There will be about 10000 house-month-region observations in our dataset.

3.2 Data-preprocessing

20pct for training conduct polynomial transformation of features polynomial expansion, variable scaling

3.3 Machine Learning Algorithm

We are going to train a machine learning model on residential housing prices in Copenhagen in 2023 in each postal coded region. The feature variables include structural features such as house area and age, and external variables include location, average income of the neighbourhood and weather conditions of each city. We are going to test the model with 2022 data and data from other cities such as Aarhus.

We are going to apply the polynomial and LASSO regression on the data to obtain the coefficient of deciding factors on house price in different cities in Denmark and in different temporal sphere.

We will handle the under-fitting and over-fitting problem when applying the machine learning model to get a more precise version of the house price prediction.

We will conduct cross-validation with 10-fold method. We split the data into 10 fold even sized validation bins. For each bin fit model on the data outside the validation bin. We transform and predict the target in the validation bin. We pick the model which performed the best on the validation data during cross validation.

visualize partial effects with partial dependence plots

4 Empirical Analysis

4.1 Polynomial

A common starting point in modelling house price in the literature is polynomial linear regression. We conduct LASSO regularization penalizing parameters to tackle the problem of over- and under-fitting, and run a 10-fold cross validation to tune for the best hyperparameter λ given the small size of our dataset. The result can be found below.

/Figure of predicted values against real test data./

The best hyperparameter from cross validation is given. The RMSE against each hyperparameter tuned is given as follows.

4.1.1 Learning Curve

We examine the learning curve to find out if our model faces overfitting (high variance) or under-fitting (high bias) problem following (Raschka and Mirjalili (2019) p.201-6). The x axis is the training data size, and the y axis is the model training and validation accuracies.

4.1.2 Validation Curve

We also leverage the validation curve to addressing over- and under-fitting. The y axis is the same as the learning curve, while the x axis is the model hyperparameters

4.2 Gradient Boosting Model

Gradient boosting model is a type of ensemble methods combining different classifiers that has better generalization performance than each individual classifier. (p.223-57, Raschka and Mirjalili (2019)) As a boosting model, gradient boosting boosts weak learners to strong learners. It is different from adaptive learning with regards to how weights are updated and how weak classifiers are combined.

4.3 Model Comparison

We compare the root mean squared errors and r-squared of each model to see how well they perform in prediction.

5 Conclusion

Bibliography

Deppner, J., von Ahlefeldt-Dehn, B., Beracha, E., and Schaefers, W. (2023). Boosting the accuracy of commercial real estate appraisals: An interpretable machine learning approach. *The Journal of Real Estate Finance and Economics*.

Lin, W., Shi, Z., Wang, Y., and Yan, T. H. (2023). Unfolding beijing in a hedonic way. *Computational Economics*, 61(1):317–340.

Raschka, S. and Mirjalili, V. (2019). *Python Machine Learning - Third Edition*. Packt Publishing.

Tchuente, D. and Nyawa, S. (2022). Real estate price estimation in french cities using geocoding and machine learning. *Annals of Operations Research*, 308(1):571–608.