INCREASING TRANSPARENCY ON HOUSING MARKETS THROUGH MACHINE LEARNING

A dissertation in partial fulfillment of the requirements for the degree of

Doktor der Wirtschaftswissenschaften (Dr. rer. pol.)

submitted to the

FACULTY OF BUSINESS, ECONOMICS, MANAGEMENT INFORMATION SYSTEMS

UNIVERSITY OF REGENSBURG

submitted by

Cathrine Nagl

Master of Science in Real Estate (**IRE**|BS)

Advisors

PROF. DR. WOLFGANG SCHÄFERS

PROF. DR. TOBIAS JUST

Date of Disputation

28. October 2022

Acknowledgements

First and foremost, I would like to express my appreciation to my doctoral supervisor Prof. Dr. Wolfgang Schäfers. Thank you for your great and constant support during the last years, the inspiring discussions, your confidence in my abilities and our successfully completed projects. It was quite a pleasure. In addition, I would also like to thank my second doctoral supervisor, Prof. Dr. Tobias Just, for his willingness for supervision and the valuable comments.

I would like to express my sincere thanks to my colleagues for the enjoyable time at the chair. A special thanks goes to my co-authors Moritz Stang and Bastian Krämer. As project and research partners, we spent many hours together with inspiring and often sprawling discussions and the elaboration of ideas. You have redefined the term team spirit on a new level.

I also want to appreciate the ongoing support of my friends, which really helped me in challenging times. My special thanks go to Marie Wieck and Stella Zimmermann, for the unforgettable time during the IRE|BS Master studies and for the constant emotional support in the last years.

This project would not have been possible without my parents, Roland and Susanne. I feel beyond blessed to have you always by my side and want to express my gratitude for your outstanding care and support throughout the years. Your relentless belief in me carries me through life. I would also like to thank my sister Paula for your great support and dedicate a special mention to Flora, who has reminded me every day for the past 1.5 years that there is a life outside of work.

My deepest thanks go to my husband Dr. Maximilian Nagl. Without you, I would not have been able to finish this dissertation. *One step closer.* I thank you for your tireless care and support, your insightful suggestions, and your unconditional love. You have shown me for what and for whom achieving goals in life is worthwhile. *For a thousand years.*

Content

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Background

The housing market is characterized by various players to whom different interests can be attributed. For the population at large, housing is a basic need that must be met regardless of social class and income. Homeownership, in turn, combines basic needs and ongoing management, but due to the high capital commitment, this type of investment is not available to all households and usually requires a high level of debt over a long period. Banks and other capital players finance residential property for private investors as well as investments made by professional real estate companies and project developers. Professional and private investors pursue various investment strategies with the residential asset class, whether the pure development of properties and their resale after completion or long-term leasing. In contrast to owner-occupiers, their focus is particularly on an expected return on their investment, which is to be met by the tenants or buyers.

This mixed market environment leads to a high degree of complexity, especially due to the characteristics which shape housing markets. While commercial real estate markets are highly professionalized, the heterogeneous investor structure of the housing markets reinforces an information asymmetry, which might lead to imbalances especially in periods of high uncertainty, as described by Soo (2018). Furthermore, contrary to stocks or bonds, the asset class of real estate is defined by a high lot size and therefore a high capital commitment and illiquidity, caused by prolonged negotiation and financing processes, as stated by Cajias and Freudenreich (2018).

Moreover, real estate is a highly heterogeneous asset. Various features, such as the size, year of construction, or the number of rooms, but also the location play a major role in the valuation or pricing of the apartment or house (Su et al. (2021)). Private investors in particular find it difficult to make a realistic value assessment, especially if the property is to be acquired for their residential purposes and thus also to satisfy private needs. But professional market players, such as banks, also face special challenges when valuing such highly heterogeneous assets as houses. Involving real estate experts in the financing process is therefore imperative. Real estate appraisers deal with the objective valuation of real estate and evaluate it based on defined criteria. These experts often operate on a highly local basis, as real estate markets depend on various local effects, are subject to regional influences, and therefore cannot be considered on a national level. In summary, housing markets are characterized by different actors with varying interests, contain heterogeneous goods that are rarely if ever directly comparable in their valuation, cause high capital commitments, and are influenced by local and regional effects.

As defined by Schulte et al. (2005), transparency in the real estate market can be described by the absence of information asymmetry of market actors. Existing information asymmetry is caused by market participants' different levels of information available or accessible. A preliminary for information is the availability of data, which entails the relevant message about real estate markets. In recent decades, the real estate industry has been able to achieve great progress. Trends such as Big Data and digitization mean that real estate data is being recorded and stored on a large scale for the first time. Both, capital providers and real estate companies, have realized that the availability of data supports the understanding of real estate market mechanisms on the one hand, but also build a new business strand for FinTechs and PropTechs on the other hand, as described by Cajias (2021).

Various types of data and sources are available to analyze and understand the housing market. Digitization has ensured that this data is available and accessible, for example through free or paid databases. However, the growing availability of data also brings challenges. Large volumes of data need to be stored, processed, and analyzed at an automated level. This increases the complexity of the relationships within the data, as defined by Jin et al. (2015). From a market perspective, one can distinguish between fundamental and non-fundamental data (Shiller (2000), p. 18). The first type encompasses direct observable characteristics of the asset, such as the lot size, or measurable quantities of the economic surrounding, e.g., unemployment rate. The latter type is not directly observable, such as the opinion of investors or the broad mass. As non-fundamentals are usually derived from text data, processing this data information asks for the application of advanced algorithms to make them usable for practical and academic uses.

In recent years, the field of Artificial Intelligence (AI) and its subfield Machine Learning (ML), originally associated especially with Computer Science and Statistics, finds increasing attention in different research areas. The foundations for machine learning research were laid in the 1940s with early models, as for example the Neural Networks by McCulloch and Pitts

(1943), and experiences a first boom in the 1960s, as described by Fradkov (2020). However, the improved data quantity and quality and growing GPU power have increased the use of ML algorithms especially in the last two decades. The terms of ML and AI are used almost inflationary in a wide variety of processes. In the context of research, ML can fundamentally be described as algorithms, which extract information from a data set and bring them into a relationship or model (Patterson and Gibson (2017), p. 2). Suitable underlying models in the ML domain can be, among others, linear regression, decision trees, or neural networks. Furthermore, two learning branches are distinguishable, as described by Müller and Guido ((2016), p. 27): supervised learning algorithms with input variables and desired output values and unsupervised learning algorithms, where the output variables are unknown.

This dissertation investigates U.S. and German housing markets by using various machine learning algorithms. The aim of this work is not only to shed light on the process behind the used algorithms, but also to understand the drivers of residential real estate markets.

An important factor to increase transparency and understanding of market mechanisms is the analysis of the expectations and moods of the different market players, which is inherently difficult to measure. Therefore, Paper 1 develops a new method of a supervised learning algorithm to measure the sentiment of various market participants acting within the US property market by incorporating Artificial Neural Networks (ANN). The results show that this new sentiment index has a larger, more persistent and statistically significant impact on the housing market in contrast to common approaches. The superiority is also robust to different text corpuses, as news headlines and news abstracts are used. This new index provides an innovative way to increase market transparency.

Paper 2 provides an automation of the traditional market valuation approach by applying various filters and similarity functions and compare these results to statistical and machine methods and shows that the use of machine learning outperforms traditional methods under certain circumstances. To the best of my knowledge, this is the first study which offers such an exhaustive comparison of different Automated Valuation Methods (AVMs) using a unique nation-wide data set based on professional collected market values from German residential properties. The results indicate that the application of ML algorithms facilitates information processing and enhances objectivity of the valuation process, which increases the transparency of the housing market from a regulatory perspective as well.

In the context of paper 3, two kinds of model-agnostic explanation systems will be distinguished to analyze residential properties in the Top-7 cities of Germany. First, Permutation Feature Importance (PFI) models, which measures the relevance of a feature in a prediction. Second, feature effect models, namely Accumulated Local Effect Plots (ALE), which give insights in the direction and linearity of the effects. With a view to the housing market, it is investigated which features actually influence the market value of the properties and in which direction these effects work, i.e., whether the market values are influenced positively or negatively. Furthermore, it can be shown to what extent the effects represent non-linearity. Thus, important conclusions can be drawn about the effect mechanisms of the features and the market value of real estate.

Summarizing, this thesis is designed to shed light on novel aspects of the housing market, which is influenced in recent years by different trends. The data quality and quantity increased rapidly through Big Data, which affects fundamental and non-fundamental variables of the housing market. Furthermore, AI and ML, which pave the way for more efficient and fine-grained market analysis and property valuations and eXplainable Machine Learning (XAI), which brings explainability to the complex hidden mechanics of ML approaches. This dissertation finds a way to blur the boundaries of well-established and newly designed approaches, which is not only important for housing markets but is also transferable to other fields of research.

1.2 Course of Analysis and Research Questions

This section gives an overview over the course of analysis.

Paper 1: Sentiment Analysis within a Deep Learning Probabilistic Framework – New Evidence from Residential Real Estate in the United States

The purpose of Paper 1 is to introduce a new methodology for sentiment index construction. By applying ANNs, different wordlists are combined to extract the sentiment for the different stakeholder groups on the housing market. By that, the paper will answer the following research questions:

- Does a sentiment index constructed from media texts provides incremental explanatory power for the prediction of monthly changes in the S&P Case-Shiller Home Price Index?
- Does the combination of different wordlists by using ANNs outperform the traditional approach using a single wordlist?
- Is a machine learning approach able to represent the unique conditions of the U.S. housing market?

Paper 2: From Human Business to Machine Learning – Methods for Automating Real Estate Appraisals and their Practical Implications

Paper 2 focuses on the comparison of different methods used in an AVM. Furthermore, as data availability and quality differ in rural and urban locations, the requirements of different spatial areas in Germany are considered. The contribution of this paper is in answering the following research questions:

- Do machine learning methods outperform well-established AVM methods and should they therefore also be considered within the regulatory discussion of AVMs?
- Should AVMs rely on the use of one single approach, or should multiple models be integrated for different spatial areas?
- Does the performance of the methods depend on data availability and structure?

Paper 3: Explainable AI in a Real Estate Context – Exploring the Determinants of Residential Real Estate Values

In Paper 3 XAI methods are implemented to increase the transparency of machine learning approaches on a regional market level. This might counter their reputation as black box algorithms and might help to establish these methods in fields, which depend on the decisions of regulators, as for example the banking industry. Therefore, the following research questions are under investigation:

- Which characteristics are important for the market values of residential properties?
- To what extent are the features characterized by linearity or non-linearity? Are there differences here depending on different cities?
- Are there fundamental differences between condominiums and single-family homes?

1.3 Submissions and Conference Presentations

Submission details to journals, the publication status and the conference presentations are given, as follows.

Paper 1: Sentiment Analysis within a Deep Learning Probabilistic Framework – New Evidence from Residential Real Estate in the United States

Authors: Cathrine Nagl

Submission to Journal: Journal of Housing Research

Current status: Submitted (16.02.2022) and currently under review

Conference presentation:

This paper was presented at the 37th Annual Conference of the American Real Estate Society (ARES) (virtual) in March 2021.

The paper was awarded the "Manuscript Prize" in the category "Housing" and the "Doctoral Program Manuscript Prize" at the 37th ARES Conference.

Paper 2: From Human Business to Machine Learning – Methods for Automating Real Estate Appraisals and their Practical Implications

Authors: Moritz Stang, Bastian Krämer, Cathrine Nagl, Wolfgang Schäfers

Submission to Journal: Zeitschrift für Immobilienökonomie

Current status: Revise and resubmit (17.03.2022) and currently under review

Conference presentation:

This paper was presented at the 38th Annual Conference of the American Real Estate Society (ARES) (Bonita Springs, Florida) in April 2022 and is accepted at the 28th Annual Conference of the European Real Estate Society (ERES) in June 2022 (Milan).

Paper 3: Explainable AI in a Real Estate Context – Exploring the Determinants of Residential Real Estate Values

Authors: Bastian Krämer, Cathrine Nagl, Moritz Stang, Wolfgang Schäfers

Submission to Journal: Journal of Real Estate Research

Current status: Submitted (20.04.2022) and currently under review

Conference presentation:

This paper was presented at the 38th Annual Conference of the American Real Estate Society (ARES) (Bonita Springs, Florida) in April 2022 and is accepted at the 28th Annual Conference of the European Real Estate Society (ERES) in June 2022 (Milan).

1.4 References

Cajias, M. (2021), “Artificial intelligence and real estate - not just an evolution, a real game changer!”, *Journal of Property Investment & Finance*, Vol. 39 No. 1, pp. 15–18.

Cajias, M. and Freudenreich, P. (2018), “Exploring the determinants of liquidity with big data – market heterogeneity in German markets”, *Journal of Property Investment & Finance*, Vol. 36 No. 1, pp. 3–18.

Fradkov, A.L. (2020), “Early History of Machine Learning”, *IFAC-PapersOnLine*, Vol. 53 No. 2, pp. 1385–1390.

Jin, X., Wah, B.W., Cheng, X. and Wang, Y. (2015), “Significance and Challenges of Big Data Research”, *Big Data Research*, Vol. 2 No. 2, pp. 59–64.

McCulloch, W.S. and Pitts, W. (1943), “A logical calculus of the ideas immanent in nervous activity”, *The Bulletin of Mathematical Biophysics*, Vol. 5 No. 4, pp. 115–133.

Müller, A.C. and Guido, S. (2016), Introduction to machine learning with Python: A guide for data scientists, Python/Machine learning, First edition, O'Reilly, Beijing, Boston, Farnham, Sebastopol, Tokyo.

Patterson, J. and Gibson, A. (2017), Deep learning: A practitioner's approach, First edition, O'Reilly, Beijing, Boston, Farnham, Sebastopol, Tokyo.

Schulte, K.-W., Rottke, N. and Pitschke, C. (2005), “Transparency in the German real estate market”, *Journal of Property Investment & Finance*, Vol. 23 No. 1, pp. 90–108.

Shiller, R.J. (2000), Irrational exuberance, Princeton University Press, Princeton, NJ.

Soo, C.K. (2018), “Quantifying Sentiment with News Media across Local Housing Markets”, *The Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Su, T., Li, H. and An, Y. (2021), “A BIM and machine learning integration framework for automated property valuation”, *Journal of Building Engineering*, Vol. 44, p. 102636.

2 Sentiment Analysis within a Deep Learning Probabilistic Framework – New Evidence from Residential Real Estate in the United States

2.1 Abstract

This paper is devoted to the relationship between news sentiment and changes in housing market movements. It provides a novel and straightforward approach to account for heterogeneous expectations of market actors within a probabilistic framework utilizing machine learning. Our novel sentiment index shows a persistent and statistically significant explanatory power for the prediction of the housing market, in contrast to common dictionary approaches. This holds for news headlines and abstracts and different definitions of sentiment indices. Our results can be regarded as the first sentiment-based evidence of heterogeneous actors in the housing market and underline the importance of different expectations for measuring non-fundamental drivers.

Keywords: Sentiment Analysis, Neural Networks, Housing Market, Machine Learning

Acknowledgements

I gratefully acknowledge many useful comments and discussions of the participants at the 37th ARES Conference in March 2021. I would like to thank Simon Stevenson, Old Dominion University and the Lucas Institute for Real Estate Development and Finance at Florida Gulf Coast University for honoring this paper with the manuscript prize in the category “housing”. Furthermore, I would like to thank Jeremy Gabe, University of San Diego and the James R. Webb ARES Foundation, for honoring this paper with the 2021 doctoral program manuscript prize.

2.2 Introduction

Understanding the drivers and influences of real estate prices is a difficult task. Especially housing markets are characterized by a heterogeneous investment structure and various stakeholder groups. The list of relevant actors encompasses households as actual or potential buyers, sellers, renters, mortgage borrowers, and mortgage defaulters. It also involves lenders; landlords; homebuilders; construction materials suppliers; and investors. Beyond this list of direct participants, housing markets include indirect participants like politics, as governments are important actors by regulating housing supply, stimulating housing demand, and ensuring the availability of mortgage credit. Identifying and predicting the expectations and behavior of these actors can help to foresee bubbles in the housing market or turning points in the development of house prices. Following Shiller (2000), non-fundamental drivers can reflect or even form these expectations. Sentiment analysis quantifies the current mood and acts as a non-fundamental measure. In financial markets, sentiment analysis has been established for decades but is only gradually applied in real estate research. The key step in sentiment analysis is the construction of a "sentiment index" from a large amount of text data. So far, predefined word lists have been used to detect polarity in texts. However, these dictionaries assume homogeneous groups, which form their expectation or mood based on the same words and a common understanding of the text corpus. This raises the question of how we can include different expectations if we observe a heterogeneous structure, like in the housing market. The answer may lie in the increased computational power and advances in Natural Language Processing research that allow a better quality of sentiment results and open the door for new algorithm-based approaches.

Our study contributes to the existing literature of sentiment analysis and real estate research in three important ways. First, we extend the dictionary approach by developing an innovative methodology to combine different word lists, designed for different market participants, in a straightforward, probabilistic way. We show that this new sentiment index can reflect the special conditions of housing market research. It provides a persistent and statistically significant explanatory power for the prediction of monthly changes in the S&P Case-Shiller Home Price Index - the most widely reported index of single-family home prices in the U.S. The descriptive and empirical analyses give rise to the conjuncture that combining different views of market actors better predicts changes in housing prices. Second, to the best of our knowledge, this is the first paper that extracts the sentiment of news articles using headlines and abstracts focusing on the housing market. This increases the complexity of the current study, as abstracts are less concise and, in most cases, contain pros and cons. Third, we are the first to utilize the potentials of machine learning in housing market research to tackle the long-standing problem of combining different views in sentiment analysis. However, our approach is not restricted to real estate research but can be applied in any discipline where heterogeneous expectations are important which increases the interdisciplinary of our approach.

2.3 Literature Review

Numerous studies have applied the dictionary approach for text analysis to financial markets (see the literature survey by Kearney und Liu (2014)). Earlier studies mainly use a general word list, see, e.g., Cutler et al. (1988), Tetlock (2007) or Tetlock et al. (2008). They find that media texts affect stock price movements, which is in line with Shiller (2000). A particularly important step is introduced by Loughran und McDonald (2011), who argue that words may carry a different connotation in the financial context: specifically, certain key words considered positive in non-financial usage seemed instead to carry a negative connotation in the financial context, and vice versa. Loughran und McDonald (2011) developed their specialized word list, called LMCD, which has been applied in numerous other text analyses of financial markets (see, for example, Jegadeesh und Di Wu (2013), Liu und McConnell (2013), Chen et al. (2013) and Ferguson et al. (2015)). The LMCD dictionary largely replaced general dictionaries in financial market sentiment analysis.

Sentiment analysis is gradually applied in commercial real estate markets, probably because securitized real estate shares (such as REITs) are similar to stocks. Rochdi und Dietzel (2015) and Braun (2016) use online information demand to measure sentiment among REIT investors, while Ruscheinsky et al. (2018) and Koelbl (2020) apply dictionaries to media reports and the Management Discussion & Analysis (MD&A) section of REIT financial reports, respectively. Sentiment analysis has also been applied to direct real estate markets, including by Clayton et al. (2009), Freybote und Seagraves (2017), Freybote und Seagraves (2018), Heinig und Nanda (2018), and Beracha et al. (2019). The studies by Ruscheinsky et al. (2018) and Beracha et al. (2019) are notable for their uses of specialized word lists: Ruscheinsky et al. (2018) constructed a list for REIT investors based on the LMCD, while Beracha et al. (2019) constructed a separate one for direct real estate investors. Both studies underline the importance of tailored word lists. The application of sentiment analysis to housing markets is much more limited. Hohenstatt et al. (2011), Dietzel (2016), Walker (2014), Soo (2018), and Bork et al. (2020) have all constructed sentiment indicators based on single word lists or non-text related sentiment indexes. Marcato und Nanda (2016) explored whether existing sentiment indices are relevant in predicting house price movements and find that they convey valuable information.

Machine learning may offer a more powerful way to recognize patterns in human language and polarity in texts. Improved computational power and the availability of larger data sets have enabled sentiment analyses using more complex deep learning techniques and revealed the potential for improved classification. Several studies have applied machine learning to sentiment analysis of financial markets including Antweiler und Frank (2004), Das und Chen (2007), Li (2010), Huang et al. (2014), and Sinha (2015). The more computationally demanding deep learning techniques have also been applied in the context of stock return forecasting by Hájek et al. (2013), Liu et al. (2017), Borovkova und Dijkstra (2018), and Souma et al. (2019), among others. Fewer researchers to date have used machine learning to the analysis of real estate markets: Hausler et al. (2018) find that a sentiment index constructed by support vector machines has predictive power in real estate markets, while Braun et al. (2019) employ a sentiment indicator constructed by an Artificial Neural Network, emphasized the importance of real estate market sentiment in investment decision-making.

Summarizing, dictionaries are gradually applied in real estate research and all studies use only one single dictionary for sentiment extraction of texts. Studies by Ruscheinsky et al. (2018) and Beracha et al. (2019) show that specialized word lists are important. A clear drawback of the dictionary approach is that results dependent upon the chosen word list. If the connotation of the words is wrong, or not shared by all actors in the market, the constructed sentiment index cannot contain all relevant information. The use of multiple word lists should ameliorate this concern, but the combination of these word lists to one sentiment index is an open question in the literature. So far there is no study, which accounts for the heterogeneous market structure of housing markets within their sentiment analysis. Furthermore, the potentials of machine learning are not lifted as well. We add to the literature by synthesizing the established sentiment construction via dictionaries with the rising field of machine learning. We account for heterogeneity of market actors by using three different word lists developed for text analysis in different contexts. First, we use the Harvard-4 to reflect discourse among a general population, such as households. Second, we use the LMCD developed specifically to reflect discourse among participants in the financial markets, such as lenders and investors. Third, we use the Lexicoder word list developed by Young und Soroka (2012) to reflect political events such as elections and policy discussions. We apply a machine learning algorithm to combine these different dictionaries in a straightforward, probabilistic way. This allows the incorporation of different sentiments (or expectations) of market participants and simultaneously lifting all benefits of machine learning in the combination of these expectations.

2.4 Data

Our goal is to evaluate the incremental role of heterogeneous market sentiment in explaining house price movements in a model that also includes more commonly used indicators reflecting supply and demand forces affecting house prices.

Our sentiment index is based on approximately 25,000 headlines and abstracts of real estate-related articles published in the Wall Street Journal from January 2001 to December 2016. We rely on the Wall Street Journal as the sentiments of all relevant actors are likely to receive full expression. It is among the highest-circulation daily newspapers in the U.S. and while its subscriber base certainly cannot be described as “mass,” they are likely to include those who are particularly influential - or at least particularly engaged - among all groups active in the housing markets. The Wall Street Journal has been the source of data used in several studies of the effect of market sentiment among investors, including Tetlock (2007) and Tetlock et al. (2008). Moreover, the Wall Street Journal reports on housing and other real estate markets (especially on Wednesdays) and was the source for the sentiment indicator developed by Beracha et al. (2019). Finally, the Wall Street Journal also includes extensive coverage of political events and public policies, including housing and mortgage market regulation.

While several house price indices are available including those published by the Federal Housing Finance Agency, CoreLogic, and Freddie Mac, the S&P Case-Shiller Home Price Index (HPI) is the most widely reported source of information on house prices. Moreover, the S&P Case-Shiller HPI is the basis for contracts traded on the Chicago Mercantile Exchange. It is constructed monthly from pairs of transactions on the same properties, with controls intended to limit the influence of quality changes and data errors and to correct for heteroskedasticity associated with idiosyncratic drift in individual property values, as described in its original version by Case und Shiller (1990).

Other variables represent more direct demand or supply conditions driving house price movements. New privately-owned housing starts (source: U.S. Census Bureau) reflect actual construction activity to increase the supply of housing available, while new privately-owned housing units authorized by building permits (source: U.S. Census Bureau) reflect regulatory activity only - that is, not all building permits result in construction starts, though the two supply indicators are closely related. Residential construction costs are measured by the value of private residential construction (source: U.S. Census Bureau). The unemployment rate (source: U.S. Bureau of Labor Statistics) is considered primarily a demand-side indicator (rather than a driver of construction costs), as lack of employment - or the threat of it, as reflected in the unemployment rate - generally makes households unwilling to assume the financial commitment of a mortgage and, of course, makes it less likely that a lender would approve their mortgage application. Hereby, the 30-Year fixed-rate mortgage average in the United States (source: Freddie Mac) and the 10-years treasury constant maturity rate (source: Board of Governors of the Federal Reserve System) is included. The average interest rate on single-family mortgages determines the cost of borrowing, while the consumer price index inflation rate represents another indicator of the high-level health of the economy.

Finally, our model of the market forces driving house price movements includes three existing measures of general sentiment. There is a large selection of homebuyer sentiment indices, such as Fannie Mae's National Housing Survey or Qualia's Homebuyer Sentiment Index. In this study, however, the same selection of sentiment indices that Soo (2018) and Hausler et al. (2018) used is incorporated into the analysis to maintain comparability. The University of Michigan Consumer Sentiment Index (UM_CSI) is intended to measure sentiment among the consumers driving overall economic growth. The University of Michigan Survey of Home Buyers (UM_BuyHome) measures whether potential home buyers consider it to be a good or bad time to buy a house. And the Political Uncertainty Index (PU), constructed based on news coverage from 10 large newspapers including the Wall Street Journal, measures sentiment specifically regarding policy issues.

Table 2.1: Descriptive statistic - Economic time series

	Observations	Mean	Standard Deviation	Minimum	Maximum
Case-Shiller-Homeprice	177	0.0005	0.0013	-0.0032	0.0029
Consumer Price Index	177	0.0018	0.0029	-0.0177	0.0138
Building_Permits	177	0.0005	0.0504	-0.2195	0.1859
Housing_Starts	177	0.0031	0.0809	-0.1868	0.2465
Mortgage	177	-0.0027	0.0337	-0.1317	0.1510
Residential_Const	177	0.0020	0.0194	-0.0604	0.0598
Treasury	177	-0.0029	0.0683	-0.3148	0.2139
Unemployment	177	-0.0002	0.0276	-0.0746	0.0800
PU	177	0.0463	0.3355	-0.6010	0.1934
UM_CSI	177	0.0008	0.0505	-0.1807	0.1361
UM_BuyHome	177	0.0007	0.0470	-0.1643	0.2393

Table 2.1 displays descriptive statistics for the dependent variable - monthly percentage changes in the S&P Case-Shiller HPI - and each of the explanatory variables except the sentiment measures. Two details should be noted. First, the monthly growth of the S&P Case-Shiller HPI is modeled using the lagging three-month moving average of each dependent variable; therefore, the model is estimated from 177 observations of the dependent variable (April 2001-December 2016). Second, two lags of the dependent variable are included as explanatory variables, reflecting the considerable serial correlation in house price movements.

2.5 Methodology

Data Preprocessing

This paper extracts sentiment from news media by synthesizing the dictionary approach and machine learning, hereafter labelled as the "combined approach". Initially, the headlines and abstracts must be preprocessed to transmit them from human language into a computer-understandable data set. Thus, we follow the preprocessing steps of Tetlock (2007), including the removal of uninformative text parts, lemmatization, and vectorization. To reflect the importance of a word we use the so-called "term frequency-inverse document frequency" (tf-idf), described by Ramos (2003). The importance of a word increases proportionally to the number of occurrences of the word in the headline/abstract but is balanced by the frequency of the word in the entire sample. Hence, words that occur frequently tend to add little information, while words that occur rarely have more expressiveness.

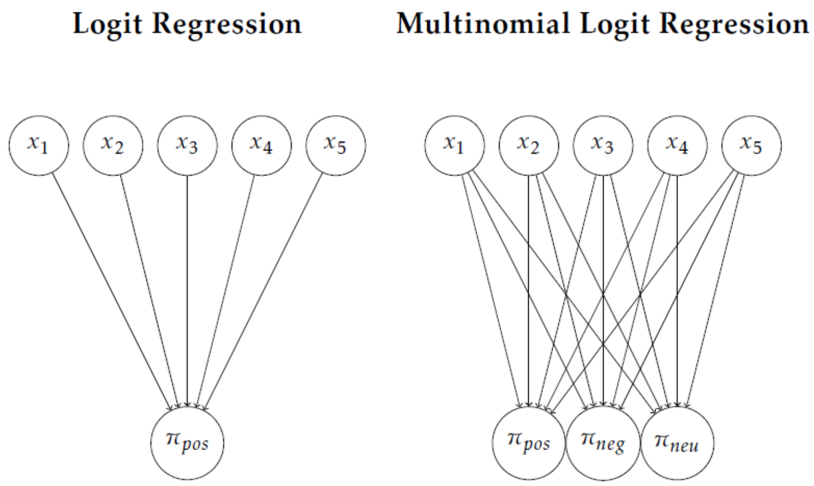
After computing the tf-idf values of each word in the text corpus, the labeling of the headlines and abstracts follows. Each text sequence is labeled individually by the three selected dictionaries. For each text sequence, the sum of tf-idf for positive and negative words, based on the selected dictionary, is calculated. According to this sum, the text sequence is labeled positive or negative. Recall that a higher tf-idf means more information. If no positive or negative word according to the dictionary appears in the text sequence or the sum of tf-idf of positive and negative words coincides, the text sequence is labeled as neutral.

The Artificial Neural Network approach in comparison

We utilize an Artificial Neural Network to determine the relation of the tf-idf values to the final sentiment label of the text corpus. For each dictionary, one Artificial Neural Network is used, trained on headlines and abstracts, which results in six different networks overall.

To introduce the structures and benefits of Artificial Neural Networks, we compare them to well-established methodologies in the literature, namely the (binary) logit and multinomial logit regression.

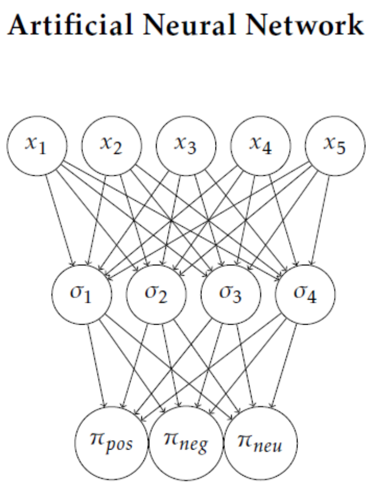
Figure 2.1: Structure of logit regression and multinomial logit regression



When facing a classification problem with two possible outcomes, binary logit regression is a common choice. For example, one can model the event of default in credit risk (defaulted vs. non-defaulted) or in the sentiment context the polarity of a text sequence (positive vs. negative), see, e.g., Ramadhan et al. (2017). As illustrated in Figure 2.1 on the left-hand side, the variables, i.e., the tf-idf values, are connected linearly and directly to the modeled probability. This approach has two drawbacks. First, we can only model positive vs. negative, neglecting the (very common) neutral polarity, and second, no joint effects, i.e., interactions, and non-linear impacts of the tf-idf values are modeled.

The first drawback can be resolved by extending the binary logit regression to a multinomial logit regression, which can model an arbitrary number of outcomes. If we compare both structures of modeling approaches in Figure 2.1, we can see that only the number of polarities is different, but the second drawback remains. To allow for any kind of non-linear and joint impacts of the tf-idf values on the estimated polarity, the multinomial logit regression is extended by an Artificial Neural Network.

Figure 2.2: Structure of Artificial Neural Networks



By incorporating (at least) one hidden layer, the Artificial Neural Network can model any kind of non-linearities and joint effects (interactions) (Buduma und Lacascio (2017)). Figure 2.2 shows that the input variables are now connected via non-linear activation functions σ in the hidden layer. This layer mimics the neurons in the human brain. Gu et al. (2020) aptly describe the operation of the neurons in the hidden layer: each neuron obtains information linearly from all input units. Then, each neuron applies a nonlinear activation function to its aggregated signal before sending its output to the next layer. In the present case, the Artificial Neural Network might be able to recognize patterns in the human language and can filter this information. Contrary, the multinomial logit regression uses only the single words without being able to make patterns or links in the text corpus under consideration.

Formally, an Artificial Neural Network consists of an input layer h_0 (i.e., the tf-idf-values of the words in the text corpus), hidden layer h_L and output layer O (i.e., the predicted probability of each polarity). The weight matrices W_l connects these layers. Neurons, parts of the hidden layers, process information via a non-linear transformation and provide these results to following neurons:

$$h_l = \sigma(h_{l-1} \cdot W_l + B_l)$$

$$O = h_L \cdot W_{L+1} + B_{L+1}$$

$l = 1, \ldots, L$ denotes the number of hidden layers. Corresponding to Rumelhart et al. (1986), the backpropagation algorithm estimates the weights and biases. As cost function, we use the categorial cross entropy, which is identical to the (negative) Likelihood of the multinomial Logit model, see Fahrmeir et al. (2013).

The output layer consists of three nodes, one for each polarity (positive, negative, neutral). Here becomes another advantage of Artificial Neural Networks apparent. While other models, such as regression trees, can only represent final classifications and Support Vector Machines only give binary classifications, the Artificial Neural Network can give probabilities of various classifications, in our case the three polarities. This enables the combination of the dictionaries under a probabilistic framework. Due to the high flexibility of the Artificial Neural Network, some hyperparameters must be pre-specified to obtain valid and reliable results and to avoid overfitting. To counteract this, we follow Gu et al. (2020). For a detailed description of the technical details, we refer to the Online Appendix.

Labeling approach with Artificial Neural Networks

The commonly used dictionary approach labels each headline or abstract independently by using the tf-idf values. Each word gets one polarity, and the final label of headline or abstract is chosen by the most frequent polarity in the text sequence. The Artificial Neural Network also uses the tf-idf values of the text sequences to find a label. The fact that all headlines or abstracts are used simultaneously during the training of the Artificial Neural Network means that the connections and similarities between these text sequences are also learned. As the Artificial Neural Networks learn patterns of human language between the different text sequences, this improves the labeling process. We use the trained Artificial Neural Network to generate combined labeling, incorporating the information of all three word lists simultaneously. A benefit of Artificial Neural Network is the softmax output vector that entails a probability distribution. In the context of our framework, the Artificial Neural Network assigns each polarity a probability. We utilize this information by constructing probability weighted labels using the Artificial Neural Networks for each dictionary. The highest (combined) probability of a label is then assigned to the headline/abstract.

To illustrate our labelling, a short example is presented below. First, the commonly known dictionary approach is used. The final label is assigned according to which dictionary-defined polarity occurs most frequently in the sentence. As can be seen in this example, the text sequence with the publishing date April 24th, 2008 is labeled differently by each of the three dictionaries: negative by Lexicoder, positive by Harvard-4, and neutral by LMCD:

Table 2.2: Labels by dictionary

Text sequence	Labels by Dictionary		
	Lexicoder	**Harvard-4**	**LMCD**
"The Brighter Side of Housing: Amid Downturn, 'Unaffordable' Is Within Reach"	Negative	Positive	Neutral

There is no straightforward way to combine the three dictionaries. The benefit of using Artificial Neural Networks is that they give probabilities for each label instead of final labels, which might implicitly assume that we are entirely sure that the dictionary assigns the true label. Hence, as the Artificial Neural Network uses the learned patterns and relations between the words and the final labels, it yields clear and important improvements. The probability results can be combined naturally to incorporate the information of several dictionaries simultaneously. The following table summarizes a sample output of the softmax function for each Artificial Neural Network trained on each dictionary:

Table 2.3: Labels under the probabilistic framework

Text sequence	Labels by Dictionary		
	Lexicoder	**Harvard-4**	**LMCD**
"The Brighter Side of Housing: Amid Downturn, 'Unaffordable' Is Within Reach"	Negative	Positive	Neutral

The weighting is conducted by computing the average of each class:

$$\pi_{combined}(positive) = \frac{19.28\% + 75.93\% + 0.01\%}{3} = 31.74\%$$

$$\pi_{combined}(neutral) = \frac{24.54\% + 5.29\% + 99.98\%}{3} = 43.27\%$$

$$\pi_{combined}(negative) = \frac{56.18\% + 18.78\% + 0.01\%}{3} = 24.99\%$$

To determine the final label, the algorithm chooses the label with the highest probability, in this case, the neutral label. Interestingly, we can also conclude that a neutral label is almost twice as likely as a negative one (43.27% versus 24.99%).

In summary, the Artificial Neural Network approach provides three improvements over common dictionary approaches. First, more information can be processed, as patterns between text sequences are incorporated. Second, instead of ultimate labels, probabilities are provided to capture the certainty of the polarity of the text sequence. Third, the probabilities enable the researchers to combine dictionaries within a probabilistic framework.

Construction of Sentiment Indices

After labeling the text sequences, we construct our sentiment indices for each dictionary and the combined approach. Following Tetlock (2007), the proportion of negative headlines in relation to the total number of text sequences builds the Pessimism Indicator *(PI)*. Tetlock (2007) finds that media content variables can forecast market activity patterns correctly. Furthermore, he examines a stronger market sensitivity to negative news. Therefore, the following concept of the PI is used:

$$PI_t = \frac{\sum_1^I negative\ text\ sequences_{i,t}}{\sum total\ number\ of\ text\ \ sequences_t}$$

The following Sentiment Quotient (SQ) is computed to test the Artificial Neural Network sentiment index's robustness in this work.

$$SQ_t = \frac{\sum_1^I positive\ text\ sequences_{i,t}}{\sum_1^I positive\ headlines_{i,t} + \sum_1^J negative\ headlines_{j,t}}$$

The Sentiment Quotient Indicator has been used by Hausler et al. (2018). This indicator excludes neutrally labeled text sequences to remove noise.

Testing the explanatory power by using a linear model

To investigate the sentiment effect of media on the housing market, a linear framework is conducted:

$$y_t = \beta_o + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

y_t is the dependent variable, in context of this study the standardized returns, i.e., subtracting the mean and divided by their standard deviation, of the S&P Case-Shiller HPI, x_{tk} the explanatory variables, β_0 the constant term, β_k the coefficient for each explanatory variable and the normally distributed error term ε_t for each observation.

To test for unit roots in the data, the Dicky-Fuller test is incorporated. Furthermore, to evaluate serial correlation and heteroscedasticity in the data, the Breusch-Godfrey and the Breusch-Pagan tests are undertaken, as well as the Jarque-Bera test for normality in the data.

2.6 Results

The Parameters of the Artificial Neural Networks

Finding a suitable architecture for the Artificial Neural Network is one of the main challenges in machine learning. For all dictionaries, the same settings are incorporated[1]. Table 2.4 illustrates the parameters during training the Artificial Neural Network, which are following the work of Gu et al. (2020). Two hidden layers, with 512 neurons in the first layer and 256

[1] The number of neurons of the two hidden layers is set to 512 and 256. The ANN is conducted with a gradient descent algorithm. The learning rate is 0.0001 and the dropout rate of 30% avoids overfitting. The batch size is set to 512. Although the model was set to train 100 epochs, early stopping allows 8-18 epochs.

neurons in the second layer, are applied. For complexity reduction and to avoid overfitting, only the 5000 most common words are used. As displayed in Table 4, the model consists of roughly 2.6 million parameters. Regarding the model accuracy, the results of the full sample show promising results between 92.8 and 94.9 percent for headlines, and 89.51 and 90.03 percent for abstracts. This is a clear indication that the model parameters are correctly chosen and meet the requirements of the underlying research question for both data sources.

Table 2.4: Training statistic for all dictionaries used

Parameters	LMCD	Harvard-4	Lexicoder
Total Parameters	2,681,859	2,681,859	2,681,859
Headlines			
Trained epochs before early stopping	10/100	8/100	8/100
In sample	0.9751	0.9511	0.9511
Out of sample	0.8691	0.8755	0.8701
Full sample	0.9433	0.9284	0.9487
Abstracts			
Trained epochs before early stopping	18/100	12/100	13/100
In sample	0.9210	0.9311	0.9305
Out of sample	0.8345	0.8415	0.8400
Full sample	0.8951	0.9042	0.9034

Sentiment Indices

As the sentiment indices should reflect the expectations and beliefs of the housing market participants, which are not captured by fundamental values, it is inevitable to meet and bear the special requirements of the market under investigation. As already discussed, the housing market contains various market actors with different expectations and information access. The Artificial Neural Network approach is conducted to combine all these different actor sentiments in one specific index.

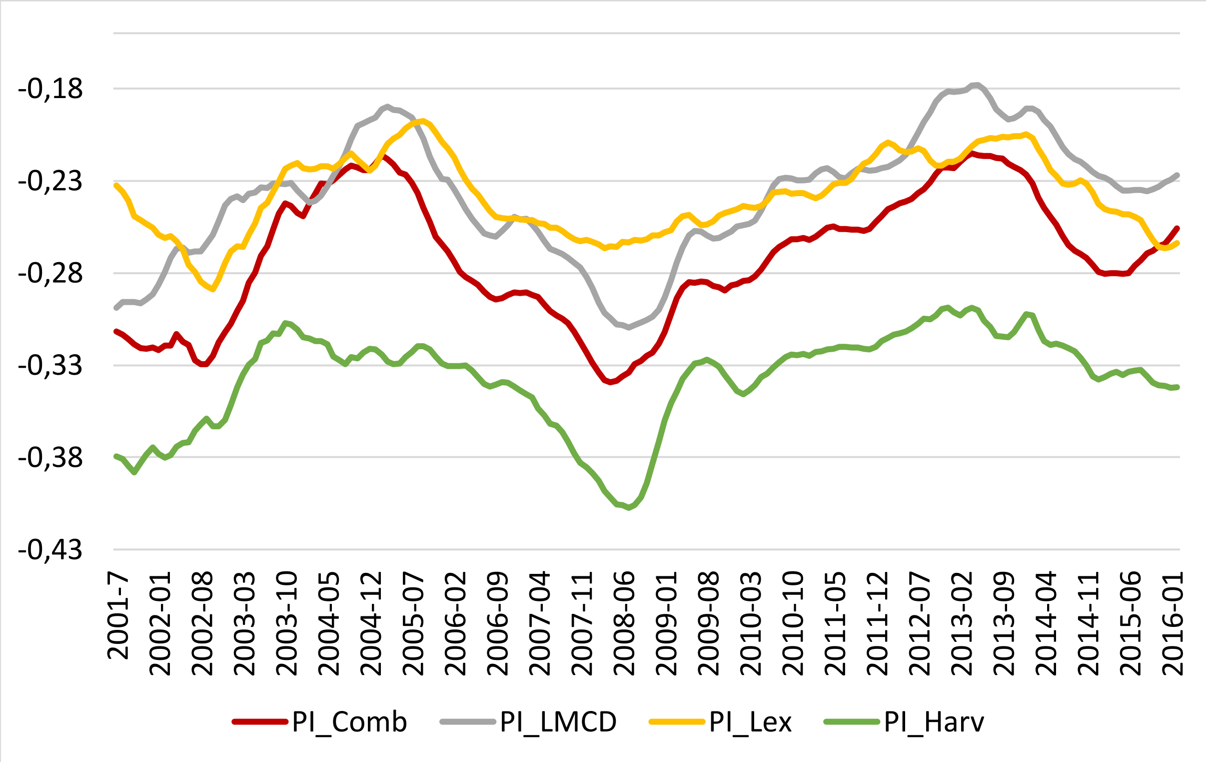
Table 2.5 represents the descriptive statistics of the four constructed sentiment indices of the Pessimism Indicator. Three different sentiment indices are implemented by using the three single dictionaries (PI_Harv, PI_LMCD, and PI_Lex) and one index following the combined approach using all three dictionaries combined by the Artificial Neural Network (PI_Comb). The PI_LMCD for abstracts indicates the highest mean and standard deviation. This might be a sign of the lack of aptitude of the LMCD dictionary for capturing the sentiment appropriately, as its fluctuation outperforms the others. However, this only applies to the abstracts of LMCD. The PI_Comb shows the lowest mean in comparison to the other indicators and the second-lowest standard deviation. Regarding the headlines, the PI_Harv shows the highest standard deviation, but not as clearly as the PI_LMCD for the abstracts.

Table 2.5: Descriptive statistics of the sentiment indices for headlines and abstracts

Abstracts	Observations	mean	std	min	max
PI_Harv	177	0.2892	0.0564	0.1058	0.4314
PI_LMCD	177	0.4662	0.1480	0.1429	0.7717
PI_Lex	177	0.2847	0.0677	0.1111	0.5085
PI_Comb	177	0.1600	0.0613	0.0000	0.3333
Headlines	Observations	mean	std	min	max
PI_Harv	177	0.3173	0.0742	0.0794	0.5106
PI_LMCD	177	0.2250	0.0645	0.0290	0.4252
PI_Lex	177	0.2242	0.0556	0.0794	0.4468
PI_Comb	177	0.2321	0.0688	0.0159	0.3966

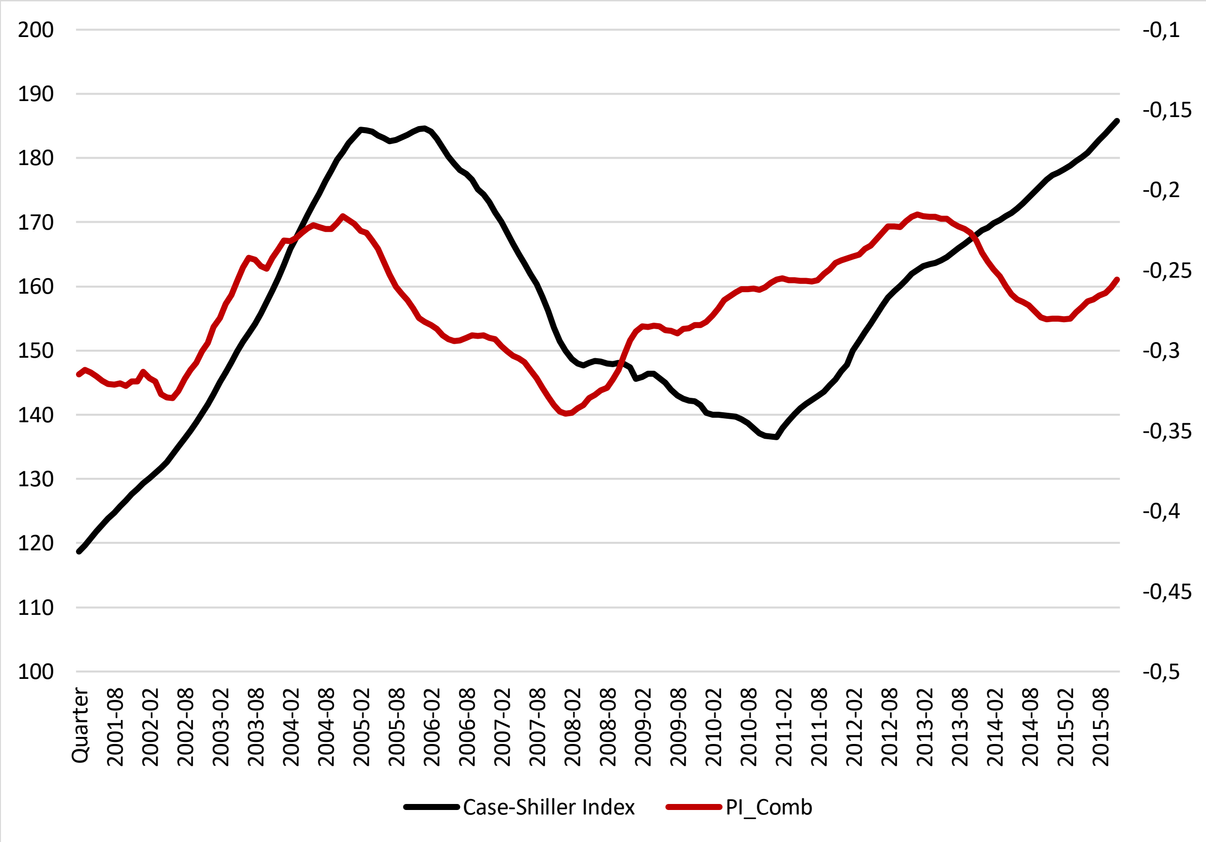
To illustrate the trends in comparison, all sentiment indices are shown in Figure 2.3. We multiply the indices with -1 to allow for a more direct interpretation as the Pessimism Indicators should be negatively related to housing prices. Comparing these sentiment indices, it is noticeable that the combined approach peaks earlier and crosses the minimum earlier than the PI_LMCD. This observation may be a first indication that the combined approach, on the one hand, recognizes and processes more information relevant to the market and, on the other hand, reacts earlier and more sensitively. Both PI_Lex and PI_Harv show a much smoother curve. However, they might not be able to anticipate market information.

Figure 2.3: Sentiment indices in comparison



Following the explanations of Soo (2018), the media sentiment of the combined approach (PI_Comb) and the S&P Case-Shiller HPI are plotted in Figure 2.4. The trend indicates that the combined index has some lead time. While the media sentiment reaches the peak at the end of 2005, the house prices are peaking at the beginning of 2007, a time lag of about one year. The media sentiment is rising at the beginning of 2009, nearly 36 months earlier than the house prices. These lead-lag patterns are in the same proportion as the results of Soo (2018). The author argues that the long transaction process and the frictions in the housing market also cause a time lag. It takes a while until the expectations and beliefs of the market participants - captured by the sentiment index in the media - can be visibly implemented in the market, as the housing market is sluggish, especially compared to the stock market. In summary, the hypothesis that the machine learning approach generates more information by combining the dictionaries seems to be supported by the descriptive analysis and will be empirically investigated in the further course of this study.

Figure 2.4: The combined sentiment index and the house price changes



The graphs suggest that sentiment and house prices have a leading effect. To test this indication not only based on descriptive analyses, we also run a regression analysis in the following. In Table 2.6 the regression output of the Pessimism Indicator for the headlines of the Wall Street Journal is displayed. The depended variable is the monthly return of the S&P Case-Shiller HPI. To control for the effects on the housing market related to fundamental variables, a set of macroeconomic variables is included with a selection of variables based on Soo (2018). All variables integrated into the regression are standardized by subtracting their mean and dividing by their standard deviation. The lag structure is optimized concerning the lowest Akaike Information Criterion (AIC). The lower the Akaike Information Criterion, the higher the information content. To obtain the optimal lag structure, all combinations up to 12 lags were tested for all four sentiment indices, resulting in a total of 16,384 regression runs. Only the optimal lags are included in the model, which ensures the optimal regression output concerning the lag structure of the sentiment indices. For all regressions, the lag structure of the control variables is the same. To compare the regression results correctly, standardized coefficients are used in this framework, which allows the comparison of the value of the coefficients across different regressions. For all independent variables, a three-month moving average is applied, in line with Soo (2018).

To determine which model provides the highest explanatory power for the housing market, we look first at the number and direction of lag structures and second at the level of Akaike Information Criterion.

Table 2.6: Pessimism Indicator – Regression Results Headlines

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
PI_LMCD (-1)	-0.0155			
	(0.0098)			
PI_LMCD (-2)	0.0165			
	(0.0100)			
PI_LMCD (-7)	-0.0130*			
	(0.0071)			
PI_LMCD (-9)	0.0109			
	(0.0067)			
PI_Lex (-1)		-0.0053		
		(0.0052)		
PI_Hav (-11)			-0.0075	
			(0.0062)	
PI_Comb (-1)				-0.0188*
				(0.0102)
PI_Comb (-2)				0.0431***
				(0.0132)
PI_Comb (-3)				-0.0276**
				(0.0108)
PI_Comb (-7)				-0.0160
				(0.0079)
PI_Comb (-9)				0.0155**
				(0.0075)
CS (-1)	0.7926***	0.7925***	0.7926***	0.8323***
	(0.0750)	(0.0755)	(0.0754)	(0.0741)
CS (-2)	0.0906	0.0908	0.0929	0.0530
	(0.0759)	(0.0761)	(0.0760)	(0.0741)
Macros	√	√	√	√
Adjusted R^2	0.8856	0.8840	0.8843	0.8912
R^2	0.8934	0.8899	0.8902	0.8993
AIC	-2214	-2214	-2215	-2222
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, and the Unemployment Rate. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12.

While Lexicoder and Harvard-4 have no significant lags and the LMCD index has only one, the combined approach has a total of four significant lags. This result indicates that the impact of the combined index is significantly larger, more sustained, and more robust than the other indices. It is also interesting to see which lags are significant, as this indicates how sustainable and long-term the effect of sentiment is.

Three indicators show a negative sign in the first lag, which is consistent with the interpretation of the pessimism indicator of Tetlock (2007), such as when pessimism in the sentiment rises, the house price returns should fall or monthly house price changes decrease. However, only the combined approach shows a significant lag structure in the first quarter and provides the highest negative impact, signaling that the indicator can measure a short-term reaction from the market actors. Fearing a deterioration in market conditions, those market participants who intended to sell their homes in the short or medium-term are selling now. Negative news about the development of the housing market particularly affects those who use their property as an investment, have financed it through a high debt ratio, or are about to sell it. A negative development on the housing market is less serious for owner-occupiers with a long property holding period. Consistent with Tetlock (2007), the second lag of the combined approach has a positive sign. The reaction in the first lag is balanced and a reversal occurs. This might indicate a bouncing effect, that more returns that are negative might be covered up by a second lag. Regarding the combined approach, we find in lags 3 and 7 negative effects, which can be interpreted as a persistent, negative effect over a long horizon, with a short correction in the ninth lag, which is rather small.

A similar lag structure can be found using the LMCD pessimism indicator. However, there is no significant impact in higher lags. This underlines the longer and persistent impact of the combined approach. Thus, as we find overall more significant lags by combining the dictionaries, the combined approach fits better to the specific circumstances of the housing market processed by different market participants. The superiority of the combined approach can also be supported by the analysis of the information criterion and adjusted R^2. The lowest AIC with -2222 is recorded by the combined approach, followed by the LMCD with -2215. Assuming that the set of macroeconomic variables already explains most in the model, the difference in the absolute amount of the AIC of the combined approach is indeed remarkable. As this information criterion penalizes uninformative model parameters, the results support the adoption that the new approach induces a higher level of information than single word lists, as it can replicate the actor requirements in the housing markets.

These results might have interesting implications for the housing market. As described by Shiller (2000), the news media are suspected to be able to influence market movements. On the one hand, the media reflect the opinion and mindset of the masses, but on the other hand, they also shape them. The media has a strong interest in market events to attract attention through headlines and thus generate higher circulation. Complex relationships in the market are simplified to make this information accessible to the general public. The results of the sentiment index suggest that this could influence the course of the market and, thus, even on the real estate cycle. In extreme cases, the media might be supportive in the formation and bursting of price bubbles. In their definition, the speculative expectation of many investors who acquire real estate to bet on a rapid price increase plays a role in particular. The identification of price bubbles is even more difficult due to the heterogeneity of players in the housing market and the economic crisis of 2008 demonstrated the impact of the U.S. housing market on the global economy. This makes it even more important to construct a precisely fitting sentiment index that can reflect the different moods of various actors and putting them into perspective.

To test whether and how the impact of the sentiment indicators with significant lags (LMCD and the combined approach) is persistent, we test the hypothesis that the sum of the lagged impact is different from zero. We can divide them roughly into short, medium, and long-term. For example, short could be the first quarter, medium second and third, and long term the fourth quarter.

Table 2.7: Qui Quadrat Test PI_Comb and PI_LMCD

PI_Comb		**PI_LMCD**	
Lag 1,2,3,7,9	-0.38%	Lag 1,2,7,9	0.64%
Lag 1,2,3,7	-1.93%**	Lag 1,2,7	-0.45%
Lag 1,2,3	-0.33%	Lag 1,2	0.85%
Lag 1,2	2.43%***		

Table 2.7 shows in each row, the sum of the coefficients over the indicated lags. We find no statistically significant different-from-zero impact over the full range of lags for both sentiment indicators. However, we can provide evidence that there is a persistent negative effect up to the seventh lag using the combined approach. This may indicate that there is a long-lasting impact of the combined pessimism indicator up to half a year, but this is reversed in the last quarter. As housing markets tend to be illiquid and slow-moving, non-transparent,

and characterized by information asymmetries, the reversal of negative media takes a longer time.

The final part of this section focuses on the sentiment conveyed by abstracts. For each headline evaluated in the previous analysis, the corresponding summary of the whole news article is available in the abstract. The following investigation aims to determine whether and how a more detailed summary of the news article offers more information in terms of sentiment.

Table 2.8: Pessimism Indicator – Regression Results Abstracts

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
PI_LMCD (-1)	-0.0041 (0.0050)			
PI_Lex (-5)		-0.0137 (0.0094)		
PI_Lex (-6)		0.0146 (0.0091)		
PI_Harv (-11)			0.0062 (0.0048)	
PI_Comb (-3)				0.0206* (0.0112)
PI_Comb (-4)				-0.0368** (0.0151)
PI_Comb (-5)				0.0153 (0.0110)
CS (-1)	0.7921*** (0.0756)	0.7886*** (0.0754)	0.7888*** (0.0754)	0.8142*** (0.0754)
CS (-2)	0.0994 (0.0770)	0.0967 (0.0770)	0.1097 (0.0775)	0.0697 (0.0760)
Macros	√	√	√	√
Adjusted R^2	0.8838	0.8844	0.8844	0.8860
R^2	0.8897	0.8909	0.8903	0.8932
AIC	-2214	-2214	-2215	-2216
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, and the Unemployment Rate. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12

Table 2.8 shows the regression results of sentiment indices using Wall Street Journal abstracts. Only the combined approach shows significant lags, where the third lag shows a positive and the fourth lag show a positive sign. This may indicate that the sentiment index based on abstracts is more sluggish and less persistent. As abstracts include more words than headlines, the probability of finding several polarities increases. This makes it more difficult to detect an explicit polarity and therefore the impact of this sentiment indicator is not as clear as for headlines. Nevertheless, the combination of word lists seems to be more flexible and can detect the nuances of clear sentiment in longer texts, such as abstracts. This is also supported by the lowest Akaike Information Criterion. Like the former analysis of the headlines, the sum over all lags is slightly negative at -0,09%, but not statistically different from zero.

Comparing all conducted regressions, i.e., headlines and abstracts, the two best regressions entail both the combined approach. This implies that the superiority of our approach does not depend on the corpus of which sentiment is extracted. Under the assumption that different word

lists incorporate the impact of different market participants, this is a piece of further evidence that actors in the housing market are heterogeneous and react differently to the news. Nevertheless, headlines seem to be particularly more influential. On the one hand, they are intended to attract attention, but on the other hand, many people only read this one sentence. As there are paywalls, not everyone has access to the full articles and only reads the headlines.

2.7 Robustness

To show that our results are robust, this section entails a battery of robustness tests. First, the regression analysis is conducted by using the Sentiment Quotient instead of the Pessimism Indicator. Table 2.9 shows the results for headlines and Table 2.10 for abstracts in the Appendix. For both, the combined approach provides the most significant lags, the expected signs, and the lowest AIC. Hence, our results are robust in the sense that the superiority of the Artificial Neural Network approach does not depend on the methodology used to construct the sentiment index.

Second, to control for the impact of the Global Financial Crisis in 2008, we rerun all regressions by adding a dummy variable for the crisis period. All results remain the same.

Third, we include additional control variables suggested by Soo (2018). In the Appendix, Table 2.11 displays the headlines and Table 2.12 the abstracts using the Pessimism Indicator. Tables 2.13 and 2.14 show the results for the Sentiment Quotient. The *News-Based Policy-Related Uncertainty* measure is incorporated. Furthermore, to capture general market sentiment, the *Surveys of Consumers of the University of Michigan* are incorporated as well. Consistent with the research question, the *Survey of Home Buyers of the University of Michigan* is part of the conducted model. Following Soo (2018), it controls for the sentiment on the housing market by surveying the opinions of 500 individuals about their assessment of home buy conditions. The integration of these three sentiment indices further increases the robustness of our model, as they control for more general sentiment in the market. As none of them shows a significant impact, this underlies the improving effect of analyzing sentiment from news media with specialized word lists for real estate. Furthermore, if we compare the resulting Akaike Information Criterion, we can conclude that it is increasing, underlining that these sentiment indices do not contain any further information for our analysis.

2.8 Conclusion

Sentiment indices are a much discussed and now widely accepted tool in academia to measure non-fundamentals that affect the market but are difficult to measure. Especially the research of Shiller (2000) laid the foundation for today's knowledge in behavioral finance about the important influence of sentiments of different actors in the market. This study shows that the media is a promising source for creating sentiment indices. However, it is important to choose the right approaches to meet the requirements of the market. In particular, the housing market is characterized by different actors for whom not only events in the market have different meanings, but who also interpret media texts and effects on the market differently.

This paper is, to the best of our knowledge, the first to use different dictionaries to capture the tone of news concerning the heterogeneous actors in the housing market. Furthermore, this paper uses a novel and innovative method to combine these three dictionaries using an Artificial Neural Network within a deep probabilistic framework. This method improves common single word list approaches by detecting and learning patterns between text sequences, quantifying the certainty of the ultimate labels, and enabling combining the dictionaries in a straightforward probabilistic way. The new probabilistic framework provides more information and most importantly, reveals a more persistent and statistically significant impact on the housing market. Moreover, the regression analysis shows a measurable impact on the housing market that is significant across several lags. Despite the integration of other sentiment indices, this effect remains persistent. Thus, a predictive effect of our approach can be concluded.

These results hold for both headlines and abstracts, answering the research question as to whether the combination of dictionaries and machine learning can account for the different and heterogeneous actors in the housing market. This study extends our knowledge regarding constructing an appropriate sentiment index in housing markets and follows the contributions of Soo (2018), that the sentiment index based on news media may be seen as an investor sentiment proxy and thus, and may be used to measure difficult market drivers, which are classified as non-fundamentals by Shiller (2000). This has interesting implications for market participants. Investors can use sentiment indices to anticipate possible changes in prices and to underpin their investment decisions. Politicians can use the sentiment index to predict how their housing market decisions will be received. And for all actors, the sentiment index is an important tool for identifying market sentiment at an early stage and assessing possible consequences for themselves and others. In particular, sentiment analysis might be a support to detect turning points in the real estate cycle and identify impending price bubbles. Thus, this study provides a foundation for future research consistent with the seminal work of Shiller (2000).

2.9 Appendix

Table 2.9: Sentiment Quotient – Regression Results Headlines

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
SQ_LMCD (-1)	0.0141*			
	(0.0078)			
SQ_LMCD (-2)	-0.0128*			
	(0.0077)			
SQ_Lex (-1)		-0.0029		
		(0.0038)		
SQ_Harv (-1)			0.0124*	
			(0.0075)	
SQ_Harv (-2)			-0.0106	
			(0.0082)	
SQ_Harv (-4)			0.0152**	
			(0.0075)	
SQ_Comb (-1)				0.0103
				(0.0067)
SQ_Comb (-2)				-0.0266***
				(0.0088)
SQ_Comb (-3)				0.0188***
				(0.0067)
CS -1)	0.7866***	0.7931***	0.7777***	0.8246***
	(0.0754)	(0.0756)	(0.0752)	(0.0745)
CS (-2)	0.0863	0.0960	0.0838	0.0569
	(0.0777)	(0.0766)	(0.0763)	(0.0767)
Macros	√	√	√	√
Adjusted R^2	0.8849	0.8837	0.8859	0.8888
R^2	0.8915	0.8896	0.8930	0.8957
AIC	-2215	-2214	-2215	-2220
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, and the Unemployment Rate. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12

Table 2.10: Sentiment Quotient – Regression Results Abstracts

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
SQ_LMCD (-4)	0.0071			
	(0.0049)			
SQ_Lex (-10)		-0.0059		
		(0.0038)		
SQ_Harv (-11)			-0.051	
			(0.0041)	
SQ_Comb (-3)				-0-0169**
				(0.0082)
SQ_Comb (-4)				0.0177**
				(0.0082)
CS (-1)	0.7926***	0.7799***	0.7873***	0.8030***
	(0.0752)	(0.0758)	(0.0756)	(0.0748)
CS (-2)	0.0715	0.0997	0.1126	0.0835
	(0.0769)	(0.0761)	(0.0780)	(0.0761)
Macros	√	√	√	√
Adjusted R^2	0.8847	0.8849	0.8844	0.8860
R^2	0.8906	0.8908	0.8903	0.8925
AIC	-2215	-2216	-2215	-2216
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, and the Unemployment Rate. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12

Table 2.11: Pessimism Indicator – Regression Result Headlines with Sentiment Indices

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
PI_LMCD (-1)	-0.0186*			
	(0.0101)			
PI_LMCD (-2)	0.0194*			
	(0.0105)			
PI_LMCD (-7)	-0.0149**			
	(0.0073)			
PI_LMCD (-9)	0.0111			
	(0.0068)			
PI_LEX (-1)		-0.0057		
		(0.0053)		
PI_HARV (-1)			-0.0079	
			(0.0063)	
PI_COMB (-1)				-0.0195*
				(0.0103)
PI_COMB (-2)				0.0465***
				(0.0136)
PI_COMB (-3)				-0.0299***
				(0.0109)
PI_COMB (-7)				-0.0183**
				(0.0081)
PI_COMB (-9)				0.0157**
				(0.0075)
CS (-1)	0.7787***	0.7864***	0.7844***	0.8225***
	(0.0756)	(0.0762)	(0.0762)	(0.0743)
CS (-2)	0.1057	0.1022	0.1049	0.0672
	(0.0765)	(0.0769)	(0.0769)	(0.0756)
Macros	√	√	√	√
PoliticalUncert	0.0023	0.0040	0.0031	0.0016
	(0.0044)	(0.0042)	(0.0042)	(0.0043)
CPI	0.0018	0.0005	0.0009	-0.0007
	(0.0037)	(0.0037)	(0.0037)	(0.0036)
HomeBuyer	0.0059	0.0042	0.0045	0.0074*
	(0.0041)	(0.0041)	(0.0041)	(0.0040)
Constant	0.0001	0.0001*	0.0001*	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Adjusted R^2	0.8855	0.8831	0.8834	0.8914
R^2	0.8953	0.8911	0.8914	0.9014
AIC	-22110	-2210	-2210	-2220
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, the Unemployment Rate, the Political Uncertainty Index, the University of Michigan Consumer Sentiment Index, and the University of Michigan Survey of Home Buyers. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12.

Table 2.12: Pessimism Indicator – Regression Results Abstracts with Sentiment Indices

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
PI_LMCD (-9)	-0.0040			
	(0.0050)			
PI_Lex (-1)		0.0008		
		(0.0046)		
PI_Harv (-11)			0.0062	
			(0.0048)	
PI_Comb (-3)				0.0207*
				(0.0114)
PI_Comb (-4)				-0.0385**
				(0.0151)
PI_Comb (-5)				0.0158
				(0.0111)
CS (-1)	0.7856***	0.7900***	0.7824***	0.8064***
	(0.0764)	(0.0765)	(0.0762)	(0.0760)
CS (-2)	0.1103	0.0994	0.1211	0.0817
	(0.0770)	(0.0778)	(0.0783)	(0.0766)
Macros	√	√	√	√
PoliticalUncert	0.0032	0.0036	0.0034	0.0040
	(0.0042)	(0.0042)	(0.0042)	(0.0043)
CPI	0.0004	0.0005	0.0001	0.0006
	(0.0037)	(0.0037)	(0.0037)	(0.0036)
HomeBuyer	0.0044	0.0042	0.0045	0.0051
	(0.0041)	(0.0041)	(0.0041)	(0.0040)
Constant	0.0001	0.0001	0.0000	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Adjusted R^2	0.8827	0.8823	0.8835	0.8855
R^2	0.8907	0.8903	0.8914	0.8946
AIC	-2209	-2209	-2211	-2214
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, the Unemployment Rate, the Political Uncertainty Index, the University of Michigan Consumer Sentiment Index, and the University of Michigan Survey of Home Buyers. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12.

Table 2.13: Sentiment Quotient – Regression Results Headlines with Sentiment Indices

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
SQ_LMCD (-1)	-0.0142*			
	(0.0079)			
SQ_LMCD (-2)	-0.0127			
	(0.0079)			
SQ_Lex (-1)		-0.0027		
		(0.0039)		
SQ_Harv (-1)			0.0143*	
			(0.0077)	
SQ_Harv (-2)			-0.0116	
			(0.0082)	
SQ_Harv (-4)			0.0160**	
			(0.0076)	
SQ_Comb (-1)				0.0103
				(0.0067)
SQ_Comb (-2)				-0.0262***
				(0.0089)
SQ_Comb (-3)				0.0189***
				(0.0067)
CS (-1)	0.7788***	0.7875***	0.7707***	0.8183***
	(0.0762)	(0.0763)	(0.0757)	(0.0752)
CS (-2)	0.0967	0.1063	0.0952	0.0665
	(0.0787)	(0.0774)	(0.0769)	(0.0774)
Macros	√	√	√	√
PoliticalUncert	0.0030	0.0037	0.0052	0.0028
	(0.0043)	(0.0042)	(0.0042)	(0.0042)
CPI	0.0012	0.0007	-0.0000	-0.0000
	(0.0037)	(0.0037)	(0.0037)	(0.0036)
HomeBuyer	0.0040	0.0039	0.0048	0.0044
	(0.0041)	(0.0041)	(0.0041)	(0.0040)
Constant	0.0001	0.0001	0.0001**	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Adjusted R^2	0.8839	0.8826	0.8855	0.8877
R^2	0.8925	0.8906	0.8946	0.8967
AIC	-2210	-2209	-2212	-2215
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, the Unemployment Rate, the Political Uncertainty Index, the University of Michigan Consumer Sentiment Index, and the University of Michigan Survey of Home Buyers. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12.

Table 2.14: Sentiment Quotient – Regression Results Abstracts with Sentiment Indices

	LMCD	**Lexicoder**	**Harvard-4**	**CombANN**
SQ_LMCD (-4)	0.0080			
	(0.0050)			
SQ_Lex (-10)		-0.0054		
		(0.0039)		
SQ_Harv (-11)			-0.0052	
			(0.0041)	
SQ_Comb (-3)				-0.0159*
				(0.0083)
SQ_Comb (-4)				-0.0178**
				(0.0083)
CS (-1)	0.7864***	0.7754***	0.7808***	0.7963***
	(0.0758)	(0.0766)	(0.0763)	(0.0756)
CS (-2)	0.0815	0.1086	0.1241	0.0916
	(0.0776)	(0.0769)	(0.0789)	(0.0769)
Macros	✓	✓	✓	✓
PoliticalUncert	0.0045	0.0032	0.0033	0.0032
	(0.0042)	(0.0042)	(0.0042)	(0.0042)
CPI	0.0001	0.0009	0.0001	0.0015
	(0.0037)	(0.0037)	(0.0037)	(0.0037)
HomeBuyer	0.0047	0.0035	0.0045	0.0034
	(0.0041)	(0.0041)	(0.0041)	(0.0041)
Constant	0.0001	0.0001	0.0000	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Adjusted R^2	0.8841	0.8837	0.8834	0.8848
R^2	0.8920	0.8916	0.8914	0.8934
AIC	-2212	-2211	-2211	-2212
No. observations	177	177	177	177

NOTES: This table reports results for the estimated linear model with monthly returns of the S&P Case-Shiller HPI, news-based sentiment, and a set of proxies as endogenous variables. CS reflects the serial correlation in house price movements in the first and second lag. The set of macroeconomic control variables includes Housing Permits, Housing Starts, the 30-Year Fixed-Rate Mortgage Average, the Value of Private Residential Construction, the 10-Year Treasury Constant Maturity Rate, the Unemployment Rate, the Political Uncertainty Index, the University of Michigan Consumer Sentiment Index, and the University of Michigan Survey of Home Buyers. * denotes significance at 90%, ** significance at 95%, *** significance at 99%. The sample period is 2001:M1–2016:M12.

2.10 References

Adha, M.R., Nurrohmah, S. and Abdullah, S. (2018), "Multinomial Logistic Regression and Spline Regression for Credit Risk Modelling", *Journal of Physics: Conference Series*, Vol. 1108 No. 1, p. 12019.

Antweiler, W. and Frank, M.Z. (2004), "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", *The Journal of Finance*, Vol. 59 No. 3, pp. 1259–1294.

Beracha, E., Lang, M. and Hausler, J. (2019), "On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination", *Journal of Real Estate Research*, Vol. 41 No. 4, pp. 605–637.

Borele, P., and Borikar, D.A. (2016), "An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques", *IOSR Journal of Computer Engineering (IOSR-JCE)*, Vol. 18 No. 2, pp. 64–69.

Bork, L., Møller, S.V. and Pedersen, T.Q. (2020), "A New Index of Housing Sentiment", *Management Science*, Vol. 66 No. 4, pp. 1563–1583.

Borovkova, S. and Dijkstra, M. (2018), "Deep Learning Prediction of the EURO STOXX 50 with News Sentiment", *SSRN Electronic Journal*.

Braun, J., Hausler, J. and Schäfers, W. (2019), "Artificial Intelligence, News Sentiment, and Property Market Liquidity", *Journal of Property Investment & Finance*.

Braun, N. (2016), "Google Search Volume Sentiment and its Impact on REIT Market Movements", *Journal of Property Investment & Finance*, Vol. 34 No. 3, pp. 249–262.

Buduma, N. and Lacascio, N. (2017), *Fundamentals of Deep Learning,* 1st ed., O'Reilly.

Carretta, A., Farina, V., Graziano, E.A. and Reale, M. (2013), "Does Investor Attention Influence Stock Market Activity? The Case of Spin-Off Deals", in Carretta, A. (Ed.), *Asset pricing, real estate and public finance over the crisis*, *Palgrave Macmillan studies in banking and financial institutions*, Palgrave Macmillan, Basingstoke, pp. 7–24.

Case, K.E. and Shiller, R.J. (1990), "Forecasting Prices and Excess Returns in the Housing Market", *Real Estate Economics*, Vol. 18 No. 3, pp. 253–273.

Chen, H., De, P. and Hwang, B.-H. (2013), "Customers as advisors: The role of social media in financial markets", *Working Paper*.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", *The Journal of Real Estate Finance and Economics*, Vol. 38 No. 1, pp. 5–37.

Cutler, D.M., Poterba, J.M. and Summers, L.H. (1988), "What Moves Stock Prices?", *National Bureau of Economic Research*.

Das, P., Füss, R., Hanle, B. and Russ, I.N. (2020), "The Cross-Over Effect of Irrational Sentiments in Housing, Commercial Property, and Stock Markets", *Journal of Banking & Finance*, p. 105799.

Das, S.R. and Chen, M.Y. (2007), "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web", *Management Science*, Vol. 53 No. 9, pp. 1375–1388.

Davis, A.K. and Tama-Sweet, I. (2012), "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A*: Language in Earnings Press Releases vs. MD&A", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804–837.

Dietzel, M.A. (2016), "Sentiment-based Predictions of Housing Market Turning Points with Google Trends", *International Journal of Housing Markets and Analysis*.

Engelberg, J. (2008), "Costly Information Processing: Evidence from Earnings Announcements", *SSRN Electronic Journal*.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), "Regression Models", in Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (Eds.), *Regression: Models, methods and applications*, Springer, Berlin, Heidelberg, pp. 21–72.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2008), "The Incremental Information Content of Tone Change in Management Discussion and Analysis", *NYU Working Paper*, No. 2451/27580.

Ferguson, N.J., Philip, D., Lam, H.Y.T. and Guo, J.M. (2015), "Media Content and Stock Returns The Predictive Power of Press", *Multinational Finance Journal*, Vol. 19 No. 1, pp. 1–31.

Ferris, S.P., Hao, Q. and Liao, M.-Y. (2013), "The Effect of Issuer Conservatism on IPO Pricing and Performance", *Review of Finance*, Vol. 17 No. 3, pp. 993–1027.

Freybote, J. and Seagraves, P. (2017), "Heterogeneous Investor Sentiment and Institutional Real Estate Investments", *Real Estate Economics*, Vol. 45 No. 1, pp. 154–176.

Freybote, J. and Seagraves, P. (2018), "The Impact of Investor Sentiment on Commercial Real Estate Market Liquidity", *Journal of Real Estate Research*, Vol. 40 No. 4, pp. 597–627.

Gu, S., Kelly, B. and Xiu, D. (2020), "Empirical Asset Pricing via Machine Learning", *The Review of Financial Studies*, Vol. 33 No. 5, pp. 2223–2273.

Hájek, P., Olej, V. and Myšková, R. (2013), "Forecasting Stock Prices using Sentiment Information in Annual Reports – A Neural Network and Support Vector Regression

Approach", *WSEAS Transactions on Business and Economics*, Vol. 10 No. 4, pp. 293–305.

Hausler, J., Ruscheinsky, J. and Lang, M. (2018), "News-based Sentiment Analysis in Real Estate: A Machine Learning Approach", *Journal of Real Estate Portfolio Management*, Vol. 35 No. 4, pp. 344–371.

Heinig, S. and Nanda, A. (2018), "Measuring Sentiment in Seal Estate – A Comparison Study", *Journal of Property Investment & Finance*, Vol. 36 No. 3, pp. 248–258.

Henry, E. (2006), "Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm", *Journal of Emerging Technologies in Accounting*, Vol. 3 No. 1, pp. 1–19.

Hohenstatt, R., Käsbauer, M. and Schäfers, W. (2011), ""Geco" and its Potential for Real Estate Research: Evidence from the U.S. Housing Market", *Journal of Real Estate Research*, Vol. 33 No. 4, pp. 471–506.

Huang, A.H., Zang, A.Y. and Zheng, R. (2014), "Evidence on the Information Content of Text in Analyst Reports", *The Accounting Review*, Vol. 89 No. 6, pp. 2151–2180.

Jegadeesh, N. and Di Wu (2013), "Word Power: A New Approach for Content Analysis", *Journal of Financial Economics*, Vol. 110 No. 3, pp. 712–729.

Kearney, C. and Liu, S. (2014), "Textual sentiment in Finance: A Survey of Methods and Models", *International Review of Financial Analysis*, Vol. 33, pp. 171–185.

Koelbl, M. (2020), "Is the MD&A of US REITs informative? A Textual Sentiment Study", *Journal of Property Investment & Finance*, Vol. 38 No. 3, pp. 181–201.

Li, F. (2010), "The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach: The Information Content of Corporate Filings", *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049–1102.

Liu, B., and McConnell, J.J. (2013), "The Role of The Media in Corporate Governance: Do the Media Influence Managers' Capital Allocation Decisions?", *Journal of Financial Economics*, Vol. 110 No. 1, pp. 1–17.

Liu, Y., Qin, Z., Li, P. and Wan, T. (2017), "Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis", *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*.

Loughran, T., and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *The Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Marcato, G. and Nanda, A. (2016), "Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market", *Journal of Real Estate Research*, Vol. 38 No. 2, pp. 165–203.

McCulloch, W., and Pitts, W. (1943), "A Logical Calculus of the Ideas Immanent in Nervous Activity", *Bulletin of Mathematical Biophysics*, Vol. 5 No. 4, pp. 115–133.

Ozik, G. and Sadka, R. (2012), "Media and Investment Management", *Available at SSRN 1633705*.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), "Thumbs up? Sentiment Classification using Machine Learning Techniques", *arXiv: cs/0205070*.

Ramadhan, W.P., Astri Novianty, S. and Casi Setianingsih, S. (2017), "Sentiment analysis using multinomial logistic regression", in *ICCEREC: The 2017 International Conference on Control, Electronics, Renewable Energy and Communications proceedings September 26-28, 2017, Tentrem Hotel, Yogyakarta-Indonesia*, IEEE, Piscataway, NJ.

Ramos, J. (2003), "Using TF-IDF to Determine Word Relevance in Document Queries", *Proceedings of the first instructional conference on machine learning*, Vol. 242, pp. 133–142.

Rochdi, K. and Dietzel, M. (2015), "Outperforming the benchmark: online information demand and REIT market performance", *Journal of Property Investment & Finance*, Vol. 33 No. 2, pp. 169–195.

Rokach, L. (2010), "Ensemble-based Classifiers", *Artificial Intelligence Review*, Vol. 33 No. 1-2, pp. 1–39.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986), "Learning Representations by Back-Propagating Errors", *Nature*, Vol. 323 No. 6088, pp. 533–536.

Ruscheinsky, J., Lang, M. and Schäfers, W. (2018), "Real Estate Media Sentiment through Textual Analysis", *Journal of Property Investment & Finance*, Vol. 36 No. 5, pp. 410–428.

Shiller, R.J. (2000), *Irrational exuberance*, Princeton University Press, Princeton, NJ.

Sinha, N.R. (2015), "Underreaction to News in the US Stock Market", *Quarterly Journal of Finance*, Vol. 6 No. 02, p. 1650005.

Soo, C.K. (2018), "Quantifying Sentiment with News Media across Local Housing Markets", *The Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Souma, W., Vodenska, I. and Aoyama, H. (2019), "Enhanced News Sentiment Analysis Using Deep Learning Methods", *Journal of Computational Social Science*, Vol. 2 No. 1, pp. 33–46.

Stone, P.J., Dunphy, D.C. and Smith, M.S. (1966), *The general inquirer: A computer approach to content analysis*, M.I.T. Press, Oxford, England.

Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008), "More Than Words: Quantifying Language to Measure Firms' Fundamentals", *The Journal of Finance*, Vol. 63 No. 3, pp. 1437–1467.

Twedt, B. and Rees, L. (2012), "Reading between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports", *Journal of Accounting and Public Policy*, Vol. 31 No. 1, pp. 1–21.

Walker, C.B. (2014), "Housing Booms and Media Coverage", *Applied Economics*, Vol. 46 No. 32, pp. 3954–3967.

Young, L. and Soroka, S. (2012), "Affective News: The Automated Coding of Sentiment in Political Texts", *Political Communication*, Vol. 29 No. 2, pp. 205–231.

Zhang, L., Wang, S. and Liu, B. (2018), "Deep Learning for Sentiment Analysis: A Survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 8 No. 4, e1253.

3 From Human Business to Machine Learning – Methods for Automating Real Estate Appraisals and their Practical Implications

3.1 Abstract

Until recently, in most countries, the use of Automated Valuation Models (AVMs) in the lending process was only allowed for support purposes, and not as the sole value-determining tool. However, this is currently changing, and regulators around the world are actively discussing the approval of AVMs. But the discussion is generally limited to AVMs that are based on already established methods such as an automation of the traditional sales comparison approach or linear regressions. Modern machine learning approaches are almost completely excluded from the debate. Accordingly, this study contributes to the discussion on why AVMs based on machine learning approaches should also be considered. For this purpose, an automation of the sales comparison method by using filters and similarity functions, two hedonic price functions – namely an OLS model and a GAM model, as well as a XGBoost machine learning approach, are applied to a dataset of 1.2 million residential properties across Germany. We find that the machine learning method XGBoost offers the overall best performance regarding the accuracy of estimations. Practical application shows that optimization of the established methods - OLS and GAM - is time-consuming and labor-intensive, and has significant disadvantages when being implemented on a national scale. In addition, our results show that different types of methods perform best in different regions and, thus, regulators should not only focus on one single method, but consider a multitude of them.

Keywords: Automated Valuation Models, Extreme Gradient Boosting, Housing Market, Machine Learning, Sales Comparison Method

3.2 Introduction

Although the real estate industry is often accused of lagging behind in terms of digitalization, the automation of processes is in fact being more and more actively discussed. In addition to the potential cost savings, ongoing improvements of computer resources and available data play an important role. Hence, it is now possible to raise data potential by automating daily processes. This potential can be leveraged in all areas of the real estate industry. Focusing on valuation, Automated Valuation Models (AVMs) have the power to change the appraisal process in many ways.

In the real estate industry, there are three different approaches to assessing properties, namely the cost approach, the income approach, and the sales comparison approach (see, e.g., Schulz et al. (2014) and Kim et al. (2020)). The latter assumes that the value of a property can be derived from the value of comparable properties, and is particularly well suited for automated real estate valuations. Various ways are known in science and practice to apply the sales comparison approach in the context of AVMs (see, Isakson (2002)). Besides the integration of filters and similarity functions, well-established hedonic price models and modern machine learning approaches can also be used for AVM construction (see, e.g., Pagourtzi et al. (2003) and Bogin and Shui (2020)). Furthermore, repeated sales methods are employed for AVM applications, see, e.g., Oust et al. (2020).

Currently, the use of AVMs in the lending process is only allowed for supporting purposes in most countries and not as a value-determining tool (Matysiak (2017) and Downie and Robson (2008)). Although there are now regulatory efforts to include AVMs in the lending process, this is only possible if the traceability, auditability, robustness and resilience of the inputs and outputs can be guaranteed (European Banking Authority (2020)). However, it remains unclear which of the abovementioned methods meet these requirements. While there is an ongoing debate about allowing the use of AVMs based on already established methods such as similarity functions or OLS regressions within the lending process, the application of modern machine learning methods is almost completely absent from the regulatory discussion. This is in fact due to the “black box” label of modern machine learning techniques. The decisions made by these methods are not as easy to understand as is the case for linear-based models due to more complex internal processes. However, in recent years, there have been various approaches to opening this black box; see for example by Friedman (2001), Goldstein et al.

(2015), Lundberg and Lee (2017) and Apley and Zhu (2020). Through these approaches, the requirements of the supervisory authority for tractability and audibility can be considered.

Therefore, the question arises as to whether modern machine learning algorithms should also be considered by the regulatory body. The objective of this paper is to contribute to this ongoing debate and deliver further insights, based on a unique nationwide dataset, into the optimal use of modern machine learning algorithms for AVMs from a theoretical and practical point of view. For this purpose, an automation of the sales comparison method by using filters and similarity functions, referred to as Expert Function (EXF), two hedonic price functions based on Ordinary Least Squares (OLS) and Generalized Additive Models (GAM), as well as the machine learning approach eXtreme Gradient Boost (XGBoost), are compared with each other.

We are the first to use a unique dataset of around 1.2 million market values of standard residential properties across Germany between 2014 and 2020, provided by a large German Banking Group, to test the four selected AVM approaches with respect to the question of whether the application of modern machine learning algorithms on a nationwide level is superior to the other approaches. The market values are based on appraiser valuations and can thus be assumed to be objective property values - unlike, for example, listing data.

The German real estate market is characterized by many different local markets whose development is often mutually independent. While metropolitan regions have seen a significant rise in values in recent years, property values in rural areas have stagnated in some cases. We are therefore also interested in whether there is one type of model which performs best in varying submarkets or whether there are structural differences. Due to the low population density, fewer observations are available in rural areas, which also raises the question of whether data availability has an impact on model performance and whether this has an influence on the decision to use machine learning algorithms for AVMs or not.

Hence, we contribute to the literature by addressing the following three research questions:

I. Do machine learning methods outperform well-established AVM methods like the OLS, the GAM and the EXF, and should they therefore also be considered within the regulatory discussion of AVMs?

II. Should AVMs rely on the use of one single approach, or should multiple models be integrated for different spatial areas?

III. Does the performance of the methods depend on data availability and structure?

Although AVMs represent a wide field in the literature, we are - to the best of our knowledge – the first to compare a filter- and similarity-based AVM approach, two well-established hedonic methods and a modern machine learning approach on a nation-wide level. Our results provide important insights into the practical application of AVMs and the discussion as to whether the usage of machine learning algorithms for the lending process should be allowed from a regulatory perspective.

We find that the machine learning method XGBoost offers the best performance regarding estimation accuracy. The EXF provides the highest transparency, but lower accuracy, as it tends to over-evaluate and does not allow calculation of the influences of individual property characteristics. The OLS and GAM are capable of doing so, but are most often outperformed by the XGBoost. Another advantage of the XGBoost is its high flexibility. While the optimization of the OLS and the GAM must be mainly done manually to achieve good model performance, the XGBoost automatically detects relevant patterns in the data. Therefore, this algorithm is better suited in practice to performing estimations based on large and complex datasets, such as nation-wide real estate valuations. However, our results also show that it is not advisable to focus on only one method when designing a nation-wide AVM. Although the XGBoost performs best across Germany, there are also regions where the EXF, the OLS or the GAM perform best. In this respect, the data availability within regions plays an important role and it is apparent that the strength of the machine learning approach cannot be improved in regions with limited training data. We therefore generally recommend testing several algorithms per region before making a final choice. In summary, our study shows that the use of machine learning algorithms for AVMs is beneficial in many situations and therefore, their approval should indeed be discussed by the regulatory authorities.

3.3 Literature Review

The following section provides a general overview of the existing literature in the field of AVMs. Due to the generally high attention devoted to this topic by the scientific community, numerous publications can be found dealing with AVMs.

The sales comparison approach normally uses a limited set of similar properties to evaluate the market value of a property, as described by French and Gabrielli (2018). Since the beginning of the computer assisted mass appraisal (CAMA) era, this approach has been automated by various researchers and is widely used in practice, especially in North America and the UK. Usually, the designed approaches follow a predefined process to identify the n most comparable sales properties from a set of N observations. The final estimation is then calculated by taking the mean or similarity-weighted mean of these comparable sales prices. Early adoptions of the similarity-based finding of comparable properties can be found in Underwood and Moesch (1982), Thompson and Gordon (1987), Cannaday (1989), McCluskey and Anand (1999) and Todora and Whiterell (2002). More recently, Brunauer et al. (2017) design an approach for valuations of self-used property based on the sales comparison method. Trawinski et al. (2017) examine the accuracy of two expert algorithms, using either the N-Latest transactions (LTA) or the N-Nearest similar properties (NSP), and compare their results with different data-driven regression models. Ciuna et al. (2017) create an approach to overcome the limitations of AVMs in markets with less available data, by means of measuring the similarity degree of the comparables. Kim et al. (2020) automate the sales comparison method to evaluate apartments in Korea and find that their approach outperforms machine learning methods. Larraz et al. (2021) use a computer-assisted expert algorithm and consider differences in characteristics compared to similar properties and their relative location.

As Borst and McCluskey (2007) show, the similarity-based automation of the sales comparison approach is also reflected in spatial autoregressive (SAR) models. The authors state that the automated sales comparison approach can be seen as a special case of a spatially lagged weight matrix model, and that there is also a less formal but clear relationship with geographically weighted regressions (GWR). Applications of SAR models can be found, among other, in McCluskey et al. (2013) and Schulz and Wersing (2021). Compared to the approach of similarity-based finding of comparable properties, the SAR model is a much more complex approach and is associated with a higher computing cost.

The hedonic price function is a well-established model that has been widely used in research for decades and was primary described by Rosen (1974). Hedonic price models do not start from the property to be valued, but from the existing information on any property available in the market, as outlined by Maier and Herath (2015). Accordingly, the property value comprises an aggregation of various attributes or characteristics regarding the amenities, micro/macro location and geodata. This also allows conclusions to be drawn about the influence of individual attributes on the value. Based on Ordinary Least Square Regression (OLS), various studies use this method in real estate valuation, for example Malpezzi (2003), Sirmans et al. (2005) and Schulz et al. (2014). In the most recent studies, OLS is used as a benchmark, for example by Zurada et al. (2011), Chrostek and Kopczewska (2013), Cajias et al. (2019) and Chin et al. (2020). For the interested reader, Metzner and Kindt (2018) and Mayer et al. (2019) provide a detailed literature review of OLS in real estate valuation.

One main disadvantage of the OLS is the dependence on the correctly specified form of the independent variables, as described by Mason and Quigley (1996). As an advanced regression model, the GAM can overcome this drawback, as it can model non-linear relationships. So-called splines are used to non-parametrically describe the relationship between the dependent and independent variables. The model was first introduced by Hastie and Tibshirani (1990) and is based on the Generalized Linear Model established by Nelder and Wedderburn (1972). Investigating the housing market in Los Angeles, Mason and Quigley (1996) are the first to use a GAM in a real estate context and find statistically significant advantages compared to OLS models. The greater flexibility and increased accuracy enable GAMs to gain further acceptance in real estate price estimation. Various other studies deal with the application of GAMs for real estate valuation, namely Pace (1998), Bao and Wan (2004), Bourassa et al. (2007), Bourassa et al. (2010) and Brunauer et al. (2010). For a detailed literature review, see Cajias and Ertl (2018).

Improved data availability and computational power have led to a whole new wave of machine learning methods, and their application to AVMs has become a widely discussed topic within academia. Machine learning methods are designed to identify non-linear structures. In addition to Artificial Neural Networks (ANN) and Support Vector Machines (SVM), tree-based models are most applied in the context of AVMs.

The idea of tree-based models dates back to Morgan and Sonquist (1963) and their automatic interaction detection (AID). The first decision tree algorithm was introduced by Quinlan (1979). The currently most commonly cited and used algorithm for decision trees was introduced by Breiman et al. (1984). Single decision trees are associated with the disadvantage that they easily overfit and therefore might perform worse on unseen data. To overcome this problem, ensemble learning techniques are used (Prajwala (2015)). Ensemble learning is defined as the combination of many "weak-learners" (e.g., single regression trees) to form one single "strong learner" (Sagi and Rokach (2018)). One efficient and commonly used version is the gradient boosting technique. The idea of gradient boosting originates back to Breiman (1997) and was primary introduced for regression trees by Friedman (2001). As Kok et al. (2017) describe, gradient-boosting models build many small decision trees subsequently, from residual-like measures of the previous trees and each tree is built from a random subsample of the dataset. Applied in real estate context, Ho et al. (2021) evaluate property prices in Hong Kong using gradient boosting trees and find that this approach outperforms other machine learning techniques like Support Vector Machines (SVM). Another example can be derived from Singh et al. (2020). The authors compare the result of gradient boosting machines with the results of a random forest regression and a linear regression approach for housing sale data in Ames, Iowa. Their findings confirm the superiority of the gradient boosting approach. Other examples can be found at Pace and Hayunga (2020) and Tchuente and Nyawa (2021). Based on the concept of gradient boosting, Tianqi and Guestrin (2016) implement the eXtreme Gradient Boosting (XGBoost) algorithm. The XGBoost is a computationally effective and highly efficient version of gradient boosting trees and applies a more regularized model structure, in order to control overfitting. Since its introduction it has often been used to tackle real-estate-specific problems. Kumkar et al. (2018), for example, compare four tree-based ensemble methods, namely bagging, random forest, gradient boosting and eXtreme gradient boosting, in terms of their efficiency in the appraisal of property in Mumbai, India. Their findings show that the XGBoost model performs better than to the other models. Sangani et al. (2017) compare the results of different gradient boosting specifications with a simple linear regression. Their analysis is based on a dataset of 2,985,217 parcels in three different counties of California. The XGBoost gradient boosting specification significantly outperforms the linear regression and is also able to perform better than almost all other specifications. Further applications of the XGBoost algorithm can be seen in Kok et al. (2017), Cajias et al. (2019) and Birkeland et al. (2021).

Although AVMs represents a wide field in the literature – to the best of our knowledge – there is currently no research comparing the performance of an advanced machine learning approach with both a filter- and similarity-based AVM and a well-established hedonic model on a nation-wide level. To address this gap in the literature, we design our own filter- and similarity-based AVM, named EXF, and apply two frequently used hedonic models, to compare their results against the performance of a modern machine learning algorithm. We use the XGBoost as our machine learning model. In several other studies, the XGBoost shows encouraging results and, compared to ANNs and SVMs, has the advantage that calculation is quicker and is therefore best suited for the size of our data set. For the hedonic models, we decide to use an OLS and a GAM. The OLS is considered to be the most widely used method in the field of AVMs and is commonly used as a benchmark. Therefore, its results are easy for readers to understand, interpret and classify. The GAM is a further development of the OLS, which can consider non-linearities by means of splines. The results of the GAM are therefore an important extension to those of the OLS. The GAM also demonstrates good performance in many other studies. Our comparison allows us to provide important insights with respect to the practical application of AVMs and the discussion on whether the usage of machine learning algorithms for the lending process should be allowed from a regulatory perspective or not.

3.4 Data

Our analysis is based on a data set of 1,212,546 residential properties across Germany. The data set is provided by a large German banking group and originates from valuations of standard residential real estate lending. The data was collected between 2014 and 2020. Table 3.1 shows how the observations are distributed over time. As the numbers show, there is a slight decreasing trend which is caused by market fluctuations. Especially, in 2020 due to COVID-19 restrictions, fewer valuations took place.

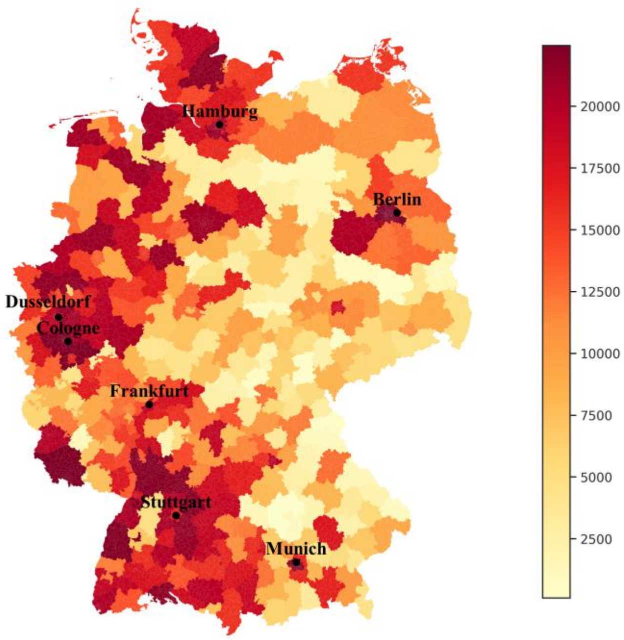
Table 3.1: Observations per year

	2014	2015	2016	2017	2018	2019	2020
n	196318	196403	176238	163365	165106	165996	149120
(%)	0.1619	0.1620	0.1453	0.1347	0.1362	0.1369	0.1230

All properties are georeferenced, making it possible to add a spatial gravity layer in order to account for spatial information. Features describing the location and neighborhood of the observations are added via Open Street Map and Acxiom[2]. The dataset was cleaned for possible outliers, erroneous values, and incompleteness.

The observations are distributed across Germany and categorized into 327 administrative districts. The division of these regions is aligned with the NUTS-3 nomenclature of the European Union. The exact distribution of individual observations can be seen in Figure 3.1. Most observations are located around the largest German metropolitan areas like Berlin, Hamburg and Munich. In addition, a difference can be observed between west and east Germany, with the east tending to have fewer observations. This is consistent with the widely diverging population figures between these regions. A comprehensive introduction to the structure of the German regions can be found at Just and Schaefer (2017), and a more detailed overview of the German real estate markets is given by Just and Maennig (2012).

Figure 3.1: Distribution of observations



The market value of the properties, based on professional appraiser valuations, is used as the target variable. In contrast to listing data, market values do not depend on subjective seller perceptions of value, but are assessed objectively by outside third parties. An overview of the average market values across the 327 administrative districts is provided in Figure 3.2. The

[2] Acxiom is an American provider of international macroeconomic and microeconomic data. Further information can be found at: https://www.acxiom.com/.

areas with the highest market values can be found in the so-called Top-7[3] cities and their commuter belts. Furthermore, the market values are by far the highest in the south of Germany and tend to be lower in the east.

Figure 3.2: Average market value

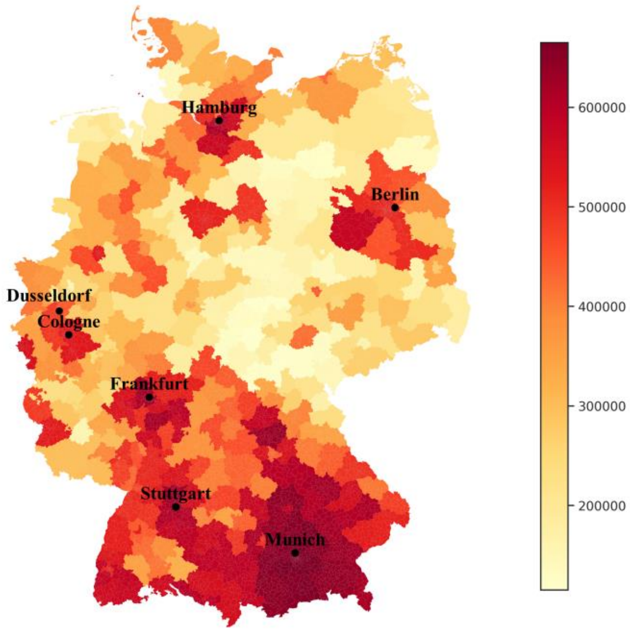


Table 3.2 shows the features included in our models and summarizes their univariate distributions. In principle, features describing the structural characteristics, micro-location and macro-location of the properties are selected. In addition, the year and quarter of the valuation is used to capture a temporal trend and seasonality. There are no correlations of concern within the data set, so that all variables can be integrated accordingly.[4]

[3] Berlin, Munich, Hamburg, Frankfurt am Main, Cologne, Dusseldorf, Stuttgart.
[4] The correlation matrix is available on request.

Table 3.2: Descriptive statistics

Variable	Unit	Mean	Median	Standard Deviation	Maximum	Minimum
Market value	Integer	228157.10	200000.00	141717.54	3860000.00	20100.00
Modernization year	Integer	1989.10	1988.00	17.19	2020.00	1950.00
Construction year	Integer	1978.48	1981.00	29.77	2020.00	1900.00
Year of valuation	Integer	2016.82	2017.00	2.00	2020.00	2014.00
Quarter of valuation	Integer	2.45	2.00	1.12	4.00	1.00
Quality grade	Integer	3.12	3.00	0.51	5.00	1.00
Macro score	Float	47.61	47.03	11.20	86.50	9.77
Micro score	Float	72.73	74.20	14.44	99.85	0.00
Living area	Float	120.31	114.68	51.69	440.00	15.00
Lot size	Float	436.48	323.00	541.66	10000.00	0.00
Latitude	Float	50.62	50.74	1.85	55.02	47.40
Longitude	Float	9.25	8.94	1.90	19.25	5.87
Basement condominium	Binary	0.38	0.00	0.48	1.00	0.00
No basement	Binary	0.19	0.00	0.39	1.00	0.00
Basement	Binary	0.44	0.00	0.50	1.00	0.00
Owner-occupied & Non-owner-occupied	Binary	0.09	0.00	0.29	1.00	0.00
Owner-occupied	Binary	0.70	1.00	0.46	1.00	0.00
Non-owner-occupied	Binary	0.21	0.00	0.41	1.00	0.00
Object subtype condominium	Binary	0.38	0.00	0.48	1.00	0.00
Object subtype detached house	Binary	0.42	0.00	0.49	1.00	0.00
Object subtype not a detached house	Binary	0.20	0.00	0.40	1.00	0.00
Condition good	Binary	0.38	0.00	0.49	1.00	0.00
Condition disastrous	Binary	0.00	0.00	0.02	1.00	0.00
Condition middle	Binary	0.45	0.00	0.50	1.00	0.00
Condition moderate	Binary	0.02	0.00	0.14	1.00	0.00
Condition bad	Binary	0.00	0.00	0.05	1.00	0.00
Condition very good	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype agglo commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype agglo cbd	Binary	0.13	0.00	0.34	1.00	0.00
Regiotype agglo middle order centre	Binary	0.13	0.00	0.34	1.00	0.00
Regiotype agglo upper order centre	Binary	0.04	0.00	0.19	1.00	0.00
Regiotype rural commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype rural middle order centre	Binary	0.07	0.00	0.26	1.00	0.00
Regiotype rural upper order centre	Binary	0.01	0.00	0.07	1.00	0.00
Regiotype urban commuter belt	Binary	0.15	0.00	0.36	1.00	0.00
Regiotype urban middle order centre	Binary	0.10	0.00	0.29	1.00	0.00
Regiotype urban upper order centre	Binary	0.07	0.00	0.26	1.00	0.00

Note: The parameter “market value” is the dependent variable in the model estimation.

Features describing the structural characteristics of the properties include the subtype of property, year of construction, modernization year, living area, lot size (only used if the property was not a condominium), use of the property, quality grade, condition and a variable denoting whether the property has a basement or not.

The subtype of a property can be either a “Condominium”, “Detached house” or “Not a detached house”. The year of modernization represents the year in which the last major refurbishment took place. The use of the building describes the possible uses, either “Owner-occupied & Non-owner-occupied”, “Owner-Occupied” or “Non-owner-occupied”. Basically, the variable describes whether a property can be rented to a third-party or not. The quality of the property is measured via a grade, on a scale ranging from 1 (very poor) to 5 (very good). The general condition of the property is represented by a categorial variable with 5 different categories ranging from disastrous to very good. The variable “Basement condominium” measures whether an apartment has an extra cellar compartment or not, whereas the “Basement” and “No Basement” variables are only valid for detached and non-detached houses. Features representing the micro-location and macro-location are latitude and longitude, different regiotypes, micro score and macro score of a location.

The regiotype was provided by Acxiom, and clusters Germany into ten different area types. In general, Acxiom defines four different spatial types: “Central-Business-District”, “Agglomeration Area”, “Urban Area” and “Rural Area”. The last three can be divided further into three sub-categories each (“Upper Centers”, “Middle Centers”, “Commuter Belt”). All addresses in Germany can be allocated to one of the ten area types. The individual area types are determined according to the respective settlement structure and population density within the municipality and its surrounding area. In most cases, the selected NUTS-3 regions can be divided further into different Regiotypes and therefore, the integration of different subtypes enables taking further local fixed effects into account.

The micro score of a location is calculated via a gravity model and reflects accessibility in the sense of proximity to selected everyday destinations. A gravity model is a common method for approximating the accessibility of a location and is based on the assumption that nearby destinations play a greater role in everyday life than more distant ones (Handy and Clifton (2001)). The score is mainly used to reduce dimensionality and complexity for the EXF. The relevant points-of-interest (POIs) are selected from the findings of Powe et al. (1995), Metzner and Kindt (2018), Yang et al. (2018), Nobis and Kuhnimhof (2018) and Huang and Dall'erba (2021) and are provided in Table 3.3. A more detailed description of the construction of the micro score of a location can be found in Appendix I.

Table 3.3: Features of the micro score of a location

Points-of-Interests	Category	Description
University	Education & Work	University campus: institute of higher education
School	Education & Work	Facility for education
Kindergarten	Education & Work	Facility for early childhood care
CBD	Education & Work	Center of the next city
Supermarket	Local Supply	Supermarket – a large store with groceries
Marketplace	Local Supply	A marketplace where goods are traded daily or weekly
Chemist	Local Supply	Shop focused on selling articles for personal hygiene, cosmetics, and household cleaning products
Bakery	Local Supply	Place for fresh bakery items
ATM	Local Supply	ATM or cash point
Hospital	Local Supply	Facility providing in-patient medical treatment
Doctors	Local Supply	Doctor's practice / surgery
Pharmacy	Local Supply	Shop where a pharmacist sells medications
Restaurant	Leisure & Food	Facility to go out to eat
Café	Leisure & Food	Place that offers casual meals and beverages
Park	Leisure & Food	A park, usually urban (municipal)
Fitness Centre	Leisure & Food	Fitness Centre, health club or gym
Movie Theater	Leisure & Food	Place where films are shown
Theater	Leisure & Food	Theatre where live performances take place
Shopping Mall	Leisure & Food	Shopping Centre– multiple stores under one roof
Department Store	Leisure & Food	Single large store selling a large variety of goods
Subway Station	Transportation	City passenger rail service
Tram Station	Transportation	City passenger rail service
Railway Station	Transportation	Railway passenger only station.
Bus Stop	Transportation	Bus stops of local bus lines.
E-Charging Station	Transportation	Charging facility for electric vehicles

Note: The descriptions of the selected Points-of-Interest is based on the explanations of Open Street Map.[5]

To account for further local fixed effects, a macro score of a location is computed. For calculation, we use a social area analysis introduced by Carpenter et al. (1955). The method assumes that a city or region can be divided into homogeneous sub-areas on the basis of different environmental variables. The variables used in our study can be seen in Table 3.4

[5] See https://wiki.openstreetmap.org/wiki/Map_features.

and are available at ZIP code level. The feature selection is based on Metzner and Kindt (2018). Further information about the macro scores can be found in Appendix II.

Table 3.4: Features for the macro score of a location

Feature	Category	Description
Educational Level	Social Status	Household structure by educational qualifications
Unemployment Rate	Social Status	Proportion of unemployed
Proportion of Children	Social Status	Proportion of population under 6 years
Purchasing Power	Economic Status	Purchasing power per household
Income Structure	Economic Status	Household structure by income
Social Security	Economic Status	Proportion of employees with social security
Relocation Behavior	Real Estate Market	Difference between inflows and outflows
Population Forecast	Real Estate Market	Population forecast for the next 5 years
Building Permits	Real Estate Market	Proportion of building permits
Construction Completions	Real Estate Market	Proportion of construction completed
Time-On-Market	Real Estate Market	Time-On-Market of properties sold

3.5 Methodology

Expert Function

The EXF uses different filters and similarity functions to determine nearby and similar comparable properties. As a result, it provides a final list of m comparables, revealing the highest degree of similarity to the property being evaluated. The next step is to estimate the market value by taking the average value of these comparables. Overall, this approach replicates the practice of traditional real estate appraisers in an automated manner. Starting with a total of N observations, a filter for spatial proximity is applied first for the EXF. Only observations within a radius of 20 km from the property to be valued are considered. Second, objects are only selected if they have the same Acxiom regiotype. Third, another filter is used to eliminate observations whose valuation date is too far in the past (< 5 years).[6] Other filters are set for the object type, occupation and presence of a basement, so as to select only

[6] For valuations longer than one year ago, an indexation with the Destatis Real Estate Price Index is applied. The index is available quarterly for five Destatis-Regiotypes starting in 2016. Mapping with the Acxiom Regiotype is performed. Further information about the index can be found at https://www-genesis.destatis.de/genesis/online.

corresponding observations. Finally, filters are set for condition and quality grade, eliminating any observations that deviate by more than one category.

After the filtering, $n \leq \mathrm{N}$ observations are left and compared with the object to be valued x^* with the aid of similarity functions. These are intended to reflect the appraiser's approach to the selection of similar properties and make it possible to select only the most similar observations for the final estimation of market value.

First, a function for spatial proximity $SP(x_i, x^*)$ is applied for all objects x_i, $i \in n$:

$$SP(x_i, x^*) = \begin{cases} 100 - 5 \cdot d(x_i, x^*), & \text{if } \; d(x_i, x^*) \in [0; 20], \\ 0, & \text{else,} \end{cases}$$

where $d(x_i, x^*)$ measures the distance between the objects as a network distance measure in kilometers (km). Next, a triangular function for measuring the similarity of the remaining features is applied:

$$tr(x_{i,f}, x_f^*, a) = \begin{cases} 100 - a(\left|x_{i,f} \; - x_f^*\right|), & \text{if} \left|x_{i,f} - x_f^*\right| < \dfrac{100}{a}, \\ 0, & \text{else,} \end{cases}$$

with $x_{i,f}$ being the value of feature f of observation i and x_f^*, the corresponding features of the object being evaluated. a describes the slope of the function. A set of different slopes was tested to find the best parameters, yielding a to be 10 for the following features: construction year, modernization year, micro score and macro score and 25 for living area and plot size.

For all objects n, we are now able to compute the feature-related similarities. These are used to calculate the overall similarity score between all x_i and x^*:

$$s(x_i, x^*) = SP(x_i, x^*) \cdot w_1 + \sum_{f=2}^{7} tr\left(x_{i,f}, x_f^*, a\right) \cdot w_f, \qquad i \in \{1, \ldots, n\},$$

with $w_1 = \frac{1}{7}$ and $w_f = \frac{1}{7}$, for all $f \in \{2, \ldots, 7\}$.

Now, we have the similarity score of the finally filtered objects n. The next step is to find the m most similar objects to $x^*, m \leq n$. Therefore, we construct a new vector v, that includes the objects in a sorted manner, so that the object with the highest overall similarity score is in the first entry and the object with the lowest overall similarity score is in the last entry. Only the first m objects of v, and therefore m most similar objects, are considered to evaluate the estimated market value of x^* by averaging their market values:

$$f(x^*) = \frac{1}{m}\sum_{i=1}^{m} f(x_i).$$

In this paper, the five most similar objects are used to estimate the market value of x^*, which is the minimum number of comparables required by law to perform a valuation by the sales comparison approach in Germany.[7]

Ordinary Least Square Regression - OLS

The first hedonic method we use is an OLS. This approach is the most commonly applied hedonic model and often used as a benchmark. Due to its simple architecture, it is easy to understand and interpret. The aim of an OLS is to explain a dependent variable y_i with independent variables $x_{i,1}, \dots, x_{i,k}$ and an error term ε_i:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i,$$

for all observations $i = 1, \dots, n$, with

$$\mu_i = E[y_i] = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}.$$

Thereby, the unknown parameters $\beta_1, \dots, \beta_k$ are estimated. The OLS assumes that the relationship between the dependent variable and independent variables is linear in parameters. Furthermore, the error terms ε_i are considered to be independent and to have a constant variance. A more detailed description can be found in Fahrmeir et al. (2013).

In order to compare the performance of the models in due course, various optimizations of the OLS are carried out. To achieve the best possible prediction power, several statistical instruments like variable transformations, interaction terms and backward stepwise regression are applied. In contrast to modern machine learning models, these optimizations must be performed manually. With 36 independent variables in the model, 630 pairwise interactions result, which must be calculated and considered for 327 different districts, summing to roughly 206,010 interactions overall. This number can easily go into the millions

[7] This procedure is based on the German guidelines for determining the mortgage lending value, see §4 BelWert.

when higher order interactions are also taken into account. This can be seen as a drawback of the OLS models.

Generalized Additive Model – GAM

The GAM is a further development of the OLS and mainly based on the concept behind the Generalized Linear Model. The relationship between the expected value of the dependent variable and the independent variables can be modelled using a monotonic link function g, like the logarithm or the inverse function. In addition, the GAM has the advantage of being able to include unspecified, non-parametric smoothing functions g_j of covariates. Consequently, we obtain the model:

$$g(\mu_i) = \beta_o + g_j\left(x_{i,1}\right) + \cdots + g_j\left(x_{i,k}\right).$$

The main advantage of the GAM compared to the OLS is its flexibility to model non-linear relationships. For the interested reader, we refer to Wood (2017).

Again, to account for locational differences, a combination of different statistical instruments like interaction terms and this time, additionally, different penalized spline types like cubic and thin plane splines have been used. Like the OLS, however, the GAM has the disadvantage that optimizations, such as the choice of spline function or interaction terms, must mainly be performed manually.

Extreme Gradient Boosting – XGBoost

EXtreme Gradient Boosting is a tree-based ensemble learning method. The idea of ensemble learning algorithms is to combine many so-called weak learners h_m, in our case, single decision trees, into one strong learner h:

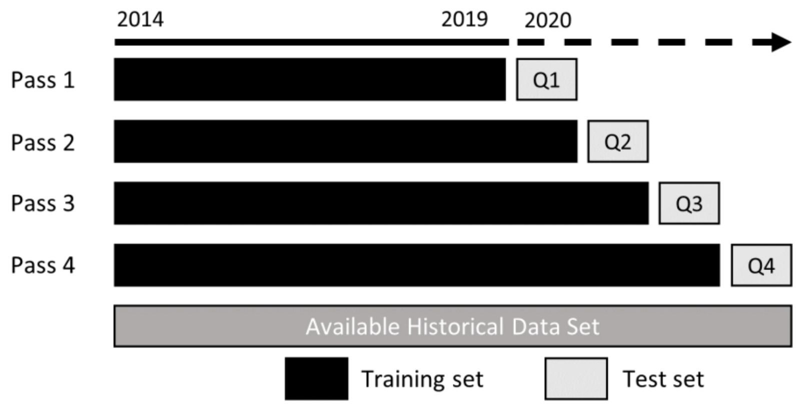
$$h(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} u_m h_m(\boldsymbol{y}|\boldsymbol{x}),$$

where u_m is used to weight the weak learners. M is the number of single trees, $\boldsymbol{x}$ is the full features space and $\boldsymbol{y}$ the response variable. Boosting is a type of ensemble learning in which the weak learners h_m are trained sequentially. Starting with one tree, the subsequent models learn from the previous errors. Gradient boosting uses the so-called gradient descent algorithm by adding new trees to minimize the loss of the model. The eXtreme Gradient Boosting is a computationally effective and highly efficient version of Gradient Boosting. In comparison to parametric and semi-parametric models, the XGBoost detects automatically complex patterns like non-linearities or higher-order interaction terms within a large amount of data, requiring for less manual optimization to account for location differences compared to the OLS and GAM. Therefore, the XGBoost convinces in practice with its flexibility, reproducibility and traceability. For more information about tree-based methods, ensemble learning and gradient boosting, the interested reader is recommended to read Hastie et al. (2001).

Testing concept

To evaluate the predictive performance of the models, an extending window approach is implemented according to Mayer et al. (2019). Figure 3.3 illustrates the testing concept.

Figure 3.3: Extending window approach



The first iteration divides the dataset into a training set with observations from Q1/2014 to Q4/2019 and a test set from Q1/2020. In the next steps, the newly available data is added to the training set, and the models are retrained and tested on data of the next quarter. The advantages of this approach are that all algorithms are tested on unseen data and thus produce unbiased, robust results. Furthermore, the testing approach provides a realistic testing scenario. In Table 3.5, the number of training and test observations for each iteration are presented.

Table 3.5: Training and test observations

Data split	Q1	Q2	Q3	Q4
Training	1,063,426	1,106,866	1,141,612	1,180,741
Test	43,440	34,746	39,129	31,805

Evaluation metrics

For each model, we compute the Mean Absolute Percentage Error (MAPE) and the Median Absolute Percentage Error (MdAPE) as accuracy measures. Unlike Mayer et al. (2019), we use the relative rather than the absolute measures of error to enable a more accurate comparison between administrative districts. Compared to the absolute measures, the relative measures provide a statement that represents the economic loss caused by the application of the algorithms much more precisely, which is very useful in our case, as we conduct a nationwide analysis involving many areas with varying levels of property market values. We thus determine the relative mean or median deviation as the economic loss or associated with by the different models. With regard to the practice of real estate financing, this loss is significant, as it shows the magnitude of misestimation in the mean or median and thus a possible risk in the event of a default of the borrower. As Rossini and Kershaw (2008) and Ecker et al. (2020) state, the MAPE and MdAPE are two precision metrics, which enable a useful comparison across different models, datasets and locations. Therefore, the MAPE is one of the most widely used metrics in the AVM literature. Examples can be found in Peterson and Flanagan (2009), Zurada et al. (2011), McCluskey et al. (2013) and Schulz et al. (2014) and Oust et al. (2020).

In order to obtain an overall picture of the strengths and weaknesses of the algorithms, we additionally provide the proportion of predictions within 10 and 20 percent (PE(x)), as well as the coefficient of determination R^2. The ratio of error buckets (PE(x)) allows us to interpret the results in a simple and intuitive way for the human brain. They show how many of the observations can be estimated within a relative deviation of 10 or 20 percent. Schulz and Wersing (2021) state that the error buckets are frequently used by practitioners when assessing valuation accuracy. A detailed description of all metrics can be found in Table 3.6.

Table 3.6: Evaluation metrics

Error	Formula	Description
Mean Absolute Percentage Error (MAPE)	$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	Mean of all absolute percentage errors. A lower MAPE signals higher prediction accuracy in percent.
Median Absolute Percentage Error (MdAPE)	$MdAPE(y,\hat{y}) = median(\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert)$	Median of all absolute percentage errors. A lower MdAPE denotes a higher precision in percent without being sensitive to outliers.
Error buckets (PE(x))	$PE(x) = 100\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert < x$	Percentage of predictions where the relative deviation is less than x%, with x being 10 and 20. A larger PE(x) signals a lower variation in the predictions.
R^2	$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Coefficient of determination. A high R^2 is an indication of better goodness of fit of the model.

3.6 Results

Results at national level for Germany

Firstly, the models are compared at a national level. In Table 3.7, the prediction errors of the entire year 2020 are summarized. For all methods, the results of the metrics evolve similarly. The more complex the structure of the approach, the better the performance. The EXF is designed to replicate the practice of traditional real estate appraisers in an automated manner and is therefore readily understandable. However, the approach provides the poorest results. Comparing these results with the performance of the OLS, often used as a baseline model, we can see a performance improvement. Relatively speaking, the MAPE of the OLS is around 18% lower and the MdAPE 19%. In addition, using an OLS results in 18% and 20% more predictions deviating less than 10 and 20 percent from their actual market value.

Analyzing the results of the GAM, we again see a boost in performance compared to the OLS. But this time the relative improvement is smaller. The MdAPE of the OLS is around 9% higher. In addition, the percentage of predictions with a relative deviation of less than 10 and 20 percent increased by 9% and 5% respectively. This might be caused by the ability of the GAM to model more complex non-linearities within the data, which is extremely difficult to manually reproduce within the OLS, and practically impossible to implement for 327 districts. This is especially so, since these manual adaptions have to be done in each of the four quarters.

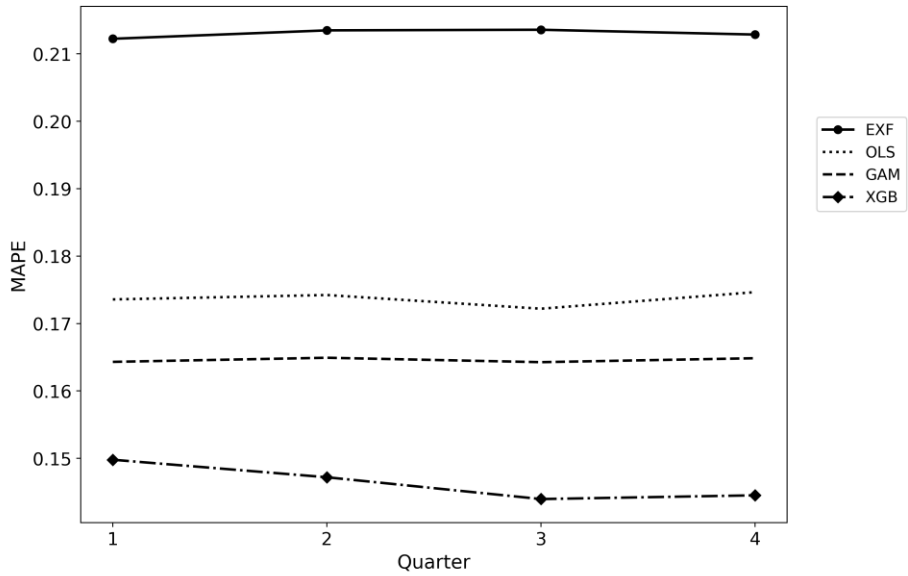
Overall, the XGBoost yields the best model performance regarding all evaluation metrics due to its ability to capture and process joint effects, non-linear relationships and high-dimensional structures within the data with comparably low manual effort. Comparing the results of the XGBoost with the EXF 43% and 33%, more observation deviate less than 10 and 20 percent from their market values.

Table 3.7: Model prediction errors 2020 throughout Germany

Models	MAPE	MdAPE	PE(10)	PE(20)	R^2
EXF	0.2130	0.1624	0.3267	0.5872	0.7735
OLS	0.1736	0.1311	0.3937	0.6940	0.8654
GAM	0.1646	0.1202	0.4273	0.7276	0.8664
XGB	0.1465	0.1084	0.4665	0.7786	0.8995

The chosen extending-window testing approach allows us to further analyze the performance of all four algorithms over the four quarters of 2020. Confirming the previous results, the line plot in Figure 3.4 shows the trends already mentioned. Additionally, it is interesting how consistently the models perform over all four quarters. Moreover, the XGBoost displays better performance the more training data it can process. The exact numbers can be seen in Appendix III.

Figure 3.4: MAPE line plot

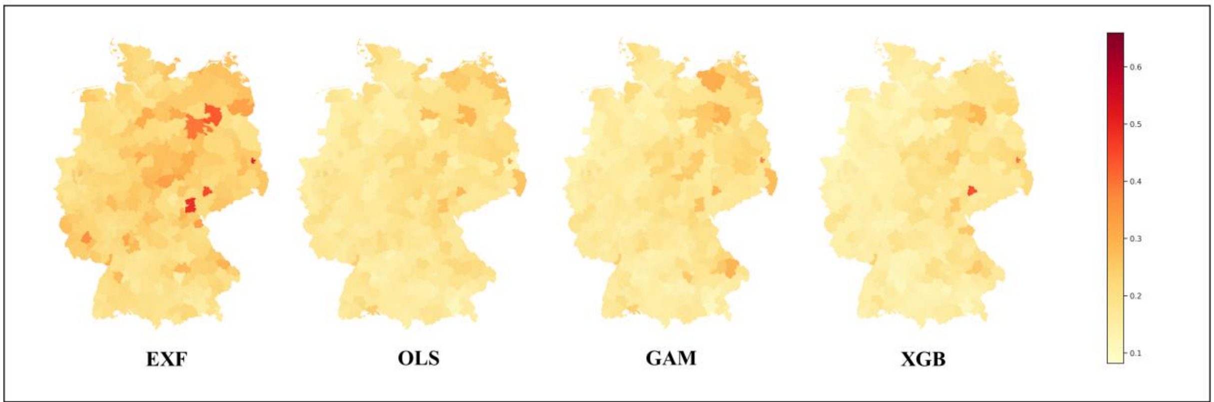


One research question of this study is to determine whether modern machine learning methods are able to outperform traditional hedonic models and the EXF approach. Analyzing our results at the national level for Germany we can clearly confirm this. The XGBoost yields a significant performance improvement compared to the EXF, OLS and GAM. This shows that in the future, regulators should also discuss the approval of machine learning methods in the field of AVMs. The application of machine learning approaches can lead to a reduction in the economic loss caused by the AVM. Machine learning algorithms are able to better assess possible risks within the lending process and can thus fulfill the actual purpose of a real estate valuation in a much more target-oriented manner.

Results at the administrative district level

After comparing the models at the national level, we want to examine the model performance in more detail. Therefore, we focus on the level of the 327 administrative districts. In Figure 3.5, the performance based on the MAPE for the different methods is shown cartographically. The maps confirm the abovementioned trends. The EXF again yields the overall poorest performance and again, it can be seen that the more complex the approach, the better the results. In addition, all four models are unsatisfactory with respect to estimating the market value in the same administrative districts. This can also be confirmed by the correlation matrices shown in Appendix III. Especially in the eastern part of Germany, the MAPE tends to be higher. This result might be caused by the lower data availability in these regions.

Figure 3.5: Error comparison at administrative district level



To obtain a better understanding of the model performance at the administrative district level, we focus on the box plots of the MAPE in Figure 3.6. Those confirm the trend displayed in Figure 3.5. The EXF again yields the overall poorest results. It delivers the largest interquartile range, the longest whiskers and contains the most outliers. The XGBoost has the lowest median MAPE of all four models, whereas it has only two extreme outliers. In contrast, the GAM and especially the OLS have a smaller range of outliers. These results indicate that the XGBoost does not always display the best model performance and therefore, different models should be used for each administrative district.

Figure 3.6: Box plots of MAPE at administrative district level

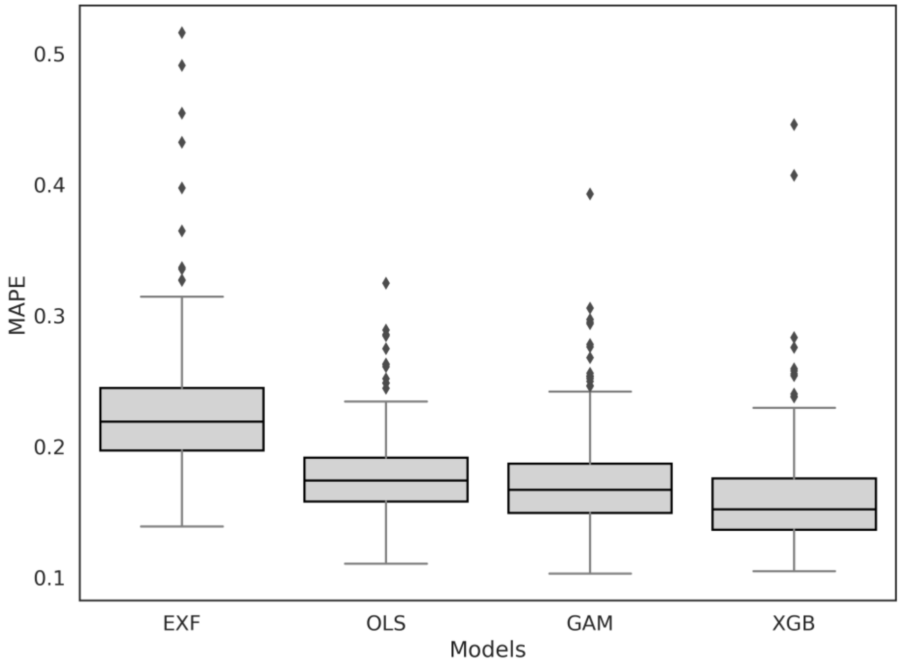


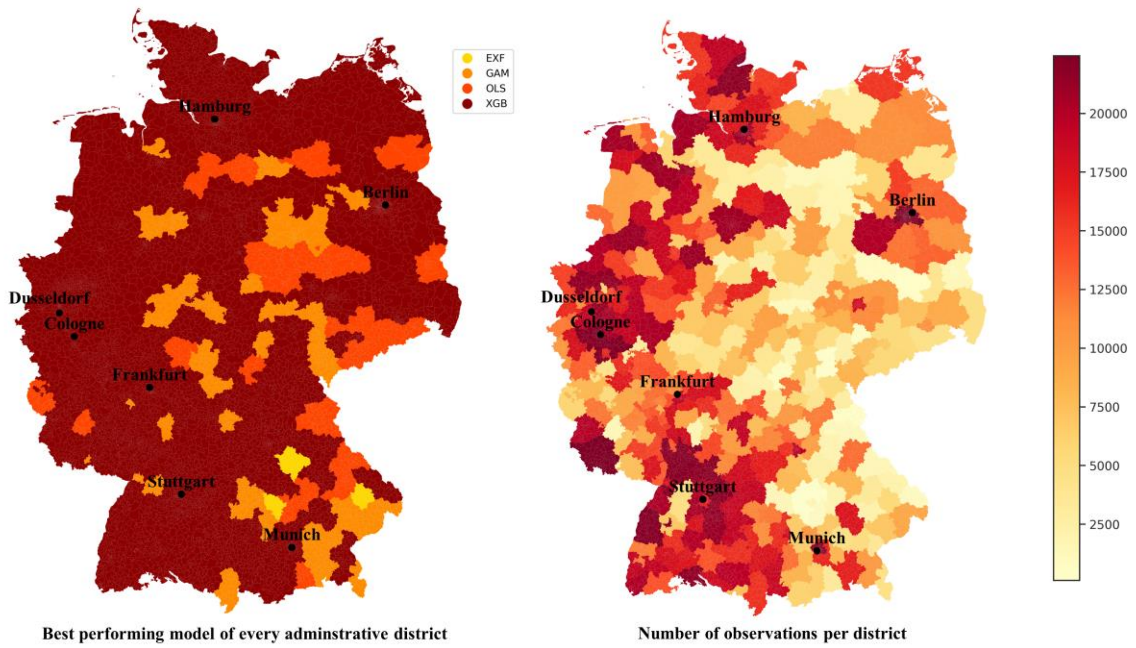
Table 3.8 shows the percentage of the administrative districts for which each model performs best. The XGBoost yields the best performance in all metrics for most administrative districts. Focusing on the hedonic approaches, the GAM and OLS are also superior in some regions, whereas EXF is the least convincing. The analysis shows that, in the case of Germany, there is no universally valid model that performs best in all administrative districts. Instead, it is advisable to apply different models in different regions.

Table 3.8: Model performance at administrative level

Models	MAPE	MdAPE	PE(10)	PE(20)	R^2
XGB	0.7920	0.7187	0.6636	0.6636	0.6422
GAM	0.1162	0.1988	0.2202	0.2202	0.0550
OLS	0.0826	0.0765	0.1101	0.1101	0.2997
EXF	0.0092	0.0061	0.0061	0.0061	0.0031

To gain a deeper understanding of the finding that different models should be used in different regions, it is useful to present the results cartographically. On the left side of Figure 3.7, the best performing model regarding the MAPE in the administrative districts is shown. On the right, the number of observations per district is presented.

Figure 3.7: Model performance and number of observations per administrative district



In the north, west and south-west of Germany, the XGBoost shows the best model performance. In contrast, especially in the south-east and east, a different picture emerges. Comparing the availability of observations with these findings, a clear dependence can be derived. In areas with many observations, the XGBoost in particular can demonstrate its strengths. By contrast, in areas with only a few observations – mostly rural regions – the GAM and OLS can also convince. Consequently, especially if one aims to implement an AVM including several different locations with a different amount of data, multiple algorithms have to be considered. By testing different algorithms, the specifics of each region can be addressed, and thus, the best model for each region can be used. This ultimately leads to a reduction of the economic loss caused by the AVM. This result shows that regulators should generally consider approving of different algorithms, and that their focus should not be on only one type of procedure.

Results at the prediction level

Lastly, we analyze the relative deviation in the market value for all four models. Accordingly, Figure 3.8 provides the density plots at the prediction level. It is evident that the EXF is negatively skewed, indicating that the approach overestimates market values to a greater extent. Converting these results into practice shows that if the EXF is used in the lending process, discounts have to be made to counteract this overvaluation. The OLS, GAM and the XGBoost are more symmetric and rather leptokurtic.

Furthermore, a cumulative distribution function plot, shown in Figure 3.9, is used to reveal whether one method outperforms another stochastically. The XGBoost is superior to the other models, with the GAM and OLS in particular being very close. In contrast, a clear gap can be seen between the OLS and the EXF. This confirms the results from above, and shows again that it is important from the regulator side also to think about approving of machine learning methods in the area of AVMs.

Figure 3.8: Density plot at prediction level

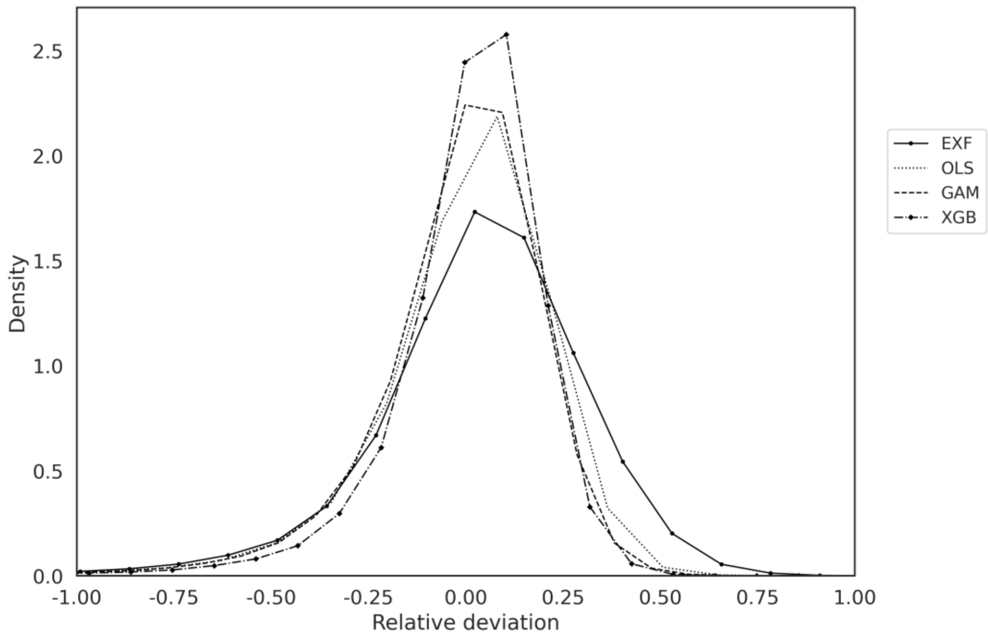
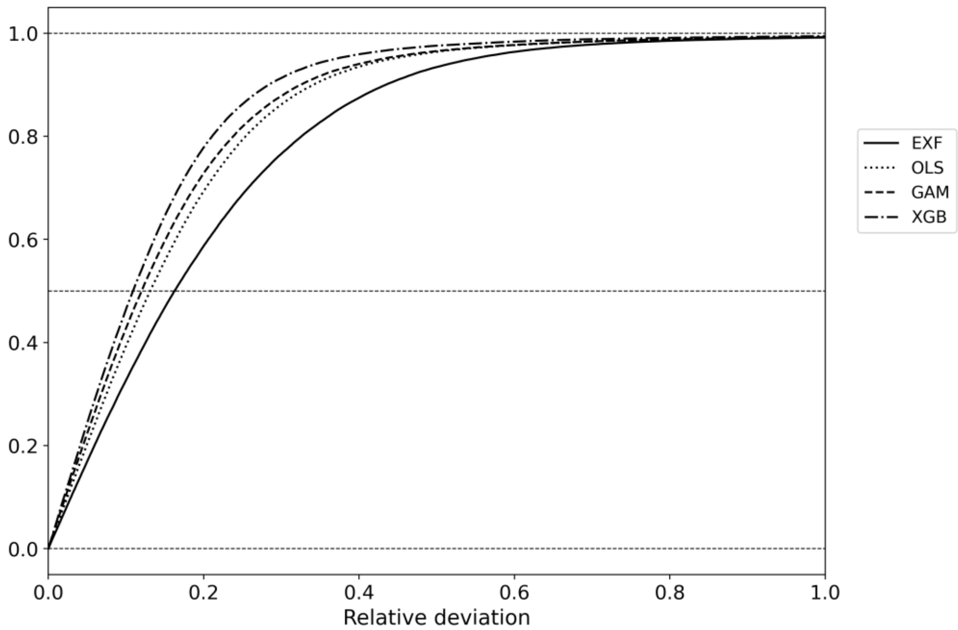


Table 3.9: Cumulative distribution function plot at prediction level



3.7 Conclusion

This study compares different approaches to constructing AVMs on a nation-wide level in order to provide empirical evidence on the regulatory debate on the future use of automated valuations. In particular, we answer the question of whether more thought should also be given to the future use of machine learning algorithms in the context of AVMs. For this purpose, an automation of the sales comparison method by using filters and similarity functions - the EXF, two hedonic price functions based on OLS and GAM, as well as the machine learning approach XGBoost, are implemented for 327 administrative districts in Germany.

As our results show, the machine learning approach XGBoost achieves the highest overall accuracy in the valuation of standard residential properties in Germany. One reason might be its ability to automatically capture and process joint effects, non-linear relationships and high-dimensional structures within a large number of observations, without requiring as many manual optimizations to account for location differences. Therefore, the XGBoost convinces in practice with its flexibility. Especially in the metropolitan areas with many observations, the relationships between the variables determining the market value seem to be much more complex, implying a need for more complex valuation models. The OLS and GAM yield weaker results. Several optimizations have been carried out to increase their predictive performance and to ensure the comparability of the models as well. However, practical application shows that the optimization of the well-established methods is time-consuming, labor-intensive and in particular, therefore shows significant disadvantages in the implementation for 327 individual districts, as it is practically infeasible. Also, the EXF does not come close to the performance of the XGBoost. The EXF even shows the weakest performance compared to the XGBoost, the OLS and the GAM. Our results indicate that the EXF tends on average to over-evaluate the predicted market values.

Furthermore, the results of our study show that for designing an AVM, there is no "one size fits all". Although the XGBoost is the best performer across the country, there are also administrative districts where the EXF, OLS, or GAM are best suited for estimating market values. In this context, it is particularly evident that the respective data availability seems to play a role. In districts with fewer observations, the traditional approaches manage to outperform the modern machine learning approach. In order to take this into account and to optimize the overall performance of AVMs, regulators should not merely allow, but actively promote the use of different types of algorithms. Before finally deploying an AVM, different types of methods should be tested for each district.

In the field of lending, a mispricing has major implications for both lenders and borrowers. Accurate model estimates are of considerable importance to ensure the resilience of the banking sector, especially in crisis periods. Our results clearly show that the approval of machine learning algorithms should be considered by regulators. We believe that machine learning algorithms have a high degree of robustness and resilience and are therefore ideally suited for AVMs. The traceability and auditability of the results required by the supervisory authorities can also be ensured by using the latest methods from the field of eXplainable Artificial Intelligence (XAI). While machine learning algorithms were considered as black box for a long time, XAI methods, like SHapely Additive exPlanations (SHAP) plots or Accumulated Local Effects (ALE) plots, are able to decode the basic decision-making process of any machine learning model. XAI is still at an early stage in the field of real estate research, but we are convinced that this will change in the coming years, and that new and important insights will be generated, which will further confirm the advantages of the use of machine learning algorithms. We therefore recommend re-examining the debate on the use of AVMs in everyday appraisals and, in particular, also including new and innovative methods.

3.8 Appendix

Appendix I – Micro Score

Our gravity model can be described using an activity function $f(A_p)$ and a distance function $f(D_{i,p})$:

$$A_{i,p} = \sum f(A_p) f(D_{i,p}).$$

$A_{i,p} \in [0{,}100]$ denotes the accessibility of point i for the POI p, whereby the activity function $f(A_p)$ specifies the relative importance of POI p, with $f(A_p) \in [0{,}1]$. $f(D_{i,p})$ measuring the travel time from point i to the POI p by using a non-symmetric sigmoidal distance function. The travel time was obtained for the selected POIs via Open Street Map and normalized using the following function:

$$L(x) = \frac{K}{(1 + Qe^{0.5x})^{\frac{1}{v}}},$$

where $K, Q \in \mathbb{R}$ and $v \in \mathbb{R}^+$ are defined for all possible distances $x \in \mathbb{R}$. Furthermore, we have:

$$K = (1 + Q)^{1+v},$$

$$Q = v \cdot \exp(B \cdot x^*),$$

$$v = \frac{\exp(B \cdot x^*) - 1}{\ln(y_i) - 1},$$

where x^* denotes a feature specific point of inflection and y^* is 0.5.

Appendix II – Macro Score

The scores $V_{j,i}(z)$ for each variable z in ZIP code i of region j are calculated using the following function:

$$V_{i,j}(z) = \left(\frac{100}{\max(z_j) - \min(z_j)}\right)(z_i - \min(z_j)),$$

where z_i denotes the value of feature z of ZIP code i. $\max(z_j)$, and $\min(z_j)$ are the maximum and minimum values of feature z in region j. As j, we define the 327 available administrative districts. Individual scores for all variables z included in the macro scores are calculated. The final macro score $MAS_{i,j}$ is computed by averaging the single scores in ZIP code i:

$$MAS_{i,j} = \frac{1}{|z|}\sum_z V_{i,j}(z).$$

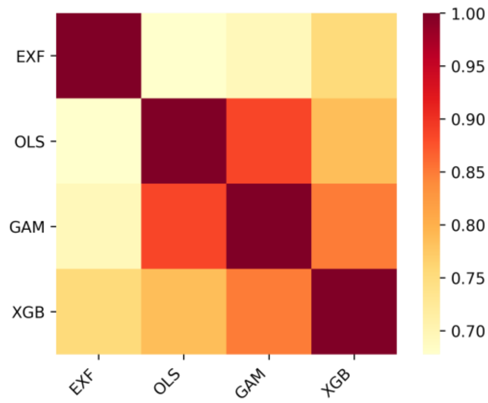
Appendix III – MAPE results on a quarterly basis

Table 3.10: MAPE on a quarterly basis throughout Germany

Models	Q1	Q2	Q3	Q4
EXF	0,2122	0,2135	0,2136	0,2129
OLS	0,1736	0,1742	0,1722	0,1747
GAM	0,1643	0,1649	0,1643	0,1649
XGB	0,1498	0,1472	0,1440	0,1445

Appendix IV – District error correlation across the models

Figure 3.9: District error correlation across the models



3.9 References

Apley, D.W. and Zhu, J. (2020), "Visualizing the effects of predictor variables in black box supervised learning models", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 82 No. 4, pp. 1059–1086.

Bao, H.X. and Wan, A.T. (2004), "On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data", *Real Estate Economics*, Vol. 32 No. 3, pp. 487–507.

Birkeland, K.B., D'silva, A.D., Füss, R. and Oust, A. (2021), "The Predictability of House Prices: "Human Against Machine", *International Real Estate Review*, Vol. 24 No. 2, pp. 139–183.

Bogin, A.N. and Shui, J. (2020), "Appraisal Accuracy and Automated Valuation Models in Rural Areas", *The Journal of Real Estate Finance and Economics*, Vol. 60 No. 1-2, pp. 40–52.

Borst, R. and McCluskey, W. (2007), "Comparative Evaluation of the Comparable Sales Method with Geostatistical Valuation Models", *Pacific Rim Property Research Journal*, Vol. 13 No. 1, pp. 106–129.

Bourassa, S.C., Cantoni, E. and Hoesli, M. (2007), "Spatial Dependence, Housing Submarkets, and House Price Prediction", *The Journal of Real Estate Finance and Economics*, Vol. 35 No. 2, pp. 143–160.

Bourassa, S.C., Cantoni, E. and Hoesli, M. (2010), "Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions", *Journal of Real Estate Research*, Vol. 32 No. 2, pp. 139–159.

Breiman, L. (1997), "Arcing the edge", Technical Report 486, Statistics Department, University of California at Berkeley.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984), "Classification and regression trees", CRC press.

Brunauer, W.A., Lang, S., Wechselberger, P. and Bienert, S. (2010), "Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna", *The Journal of Real Estate Finance and Economics*, Vol. 41 No. 4, pp. 390–411.

Brunauer, W.A., Weberndorfer, R.S. and Feilmayr, W. (2017), A statistically founded sales comparison approach, European Real Estate Society (ERES).

Cajias, M. and Ertl, S. (2018), "Spatial Effects and Non-Linearity in Hedonic Modeling*", Journal of Property Investment & Finance*, Vol. 36 No. 1, pp. 32–49.

Cajias, M., Willwersch, J. and Lorenz, F. (2019), I know where you will invest in the next year – Forecasting real estate investments with machine learning methods, European Real Estate Society (ERES).

Cannaday, R.E. (1989), "How should you estimate and provide market support for adjustments in single family appraisals", *Real Estate Appraiser and Analyst*, 55(4), 43-54.

Carpenter, D.B., Shevky, E. and Bell, W. (1955), "Social Area Analysis: Theory, Illustrative Application and Computational Procedures", *American Sociological Review*, Vol. 20 No. 4.

Chin, S., Kahn, M.E. and Moon, H.R. (2020), "Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach", *Real Estate Economics*, Vol. 48 No. 3, pp. 886–914.

Chrostek, K. and Kopczewska, K. (2013), "Spatial Prediction Models for Real Estate Market Analysis", *Ekonomia*, No. 35, pp. 25–43.

Ciuna, M., Ruggiero, M. de, Manganelli, B., Salvo, F. and Simonotti, M. (2017), "Automated Valuation Methods in Atypical Real Estate Markets Using the Mono-parametric Approach", in Gervasi, O. (Ed.), Computational science and its applications: 17th international conference, Trieste, Italy, July 3-6, 2017 proceedings, 2017, Cham, Springer, pp. 200–209.

Downie, M.-L. and Robson, G. (2008), Automated Valuation Models: An International Perspective.

Ecker, M., Isakson, H. and Kennedy, L. (2020), "An Exposition of AVM Performance Metrics", *Journal of Real Estate Practice and Education*, Vol. 22 No. 1, pp. 22–39.

European Banking Authority (2020), Final Report: Guidelines on loan origination and monitoring, Paris.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), Regression: Models, Methods and Applications, Springer-Verlag, Berlin.

French, N. and Gabrielli, L. (2018), "Pricing to market. Property valuation revisited: the hierarchy of valuation approaches, methods and models", *Journal of Property Investment & Finance*, Vol. 36 No. 4, pp. 391–396.

Friedman, J.H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, Vol. 29 No. 5, pp. 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015), "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation*", Journal of Computational and Graphical Statistics*, Vol. 24 No. 1, pp. 44–65.

Hastie, T., Friedman, J. and Tibshirani, R. (2001), *The Elements of Statistical Learning*, Springer New York, New York, NY.

Hastie, T. and Tibshirani, R. (1990), Generalized additive models, Monographs on statistics and applied probability, Vol. 43, 1. ed., Chapman and Hall, London.

Ho, W.K., Tang, B.-S. and Wong, S.W. (2021), "Predicting property prices with machine learning algorithms", *Journal of Property Research*, Vol. 38 No. 1, pp. 48–70.

Huang, Y. and Dall'erba, S. (2021), "Does Proximity to School Still Matter Once Access to Your Preferred School Zone Has Already Been Secured?", *The Journal of Real Estate Finance and Economics*, Vol. 62 No. 4, pp. 548–577.

Isakson, H. (2002), "The Linear Algebra of the Sales Comparison Approach", *Journal of Real Estate Research*, Vol. 24 No. 2, pp. 117–128.

Just, T. and Maennig, W. (Eds.) (2012), Understanding German Real Estate Markets, Springer Berlin Heidelberg, Berlin, Heidelberg.

Just, T. and Schaefer, P. (2017), "Germany's Regional Structure", in Understanding German Real Estate Markets, Springer, Cham, pp. 41–57.

Kim, Y., Choi, S. and Yi, M.Y. (2020), "Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices", *Sustainability*, Vol. 12 No. 14, p. 5679.

Kok, N., Koponen, E.-L. and Martínez-Barbosa, C.A. (2017), "Big Data in Real Estate? From Manual Appraisal to Automated Valuation", *The Journal of Portfolio Management*, Vol. 43 No. 6, pp. 202–211.

Kumkar, P., Madan, I., Kale, A., Khanvilkar, O. and Khan, A. (2018), "Comparison of Ensemble Methods for Real Estate Appraisal", in 2018 3rd International Conference on Inventive Computation Technologies (ICICT), IEEE.

Larraz, B., Alfaro-Navarro, J.-L., L Cano, E., Alfaro-Cortes, E., Garcia, N. and Gámez, M. (2021), "A computer-assisted expert algorithm for real estate valuation in Spanish cities", Environment and Planning B: Urban Analytics and City Science, Vol. 48 No. 6, pp. 1712–1727.

Lundberg, S.M. and Lee, S.-I. (2017), "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30.

Maier, G. and Herath, S. (2015), Immobilienbewertung mit hedonischen Preismodellen: Theoretische Grundlagen und praktische Anwendung, Springer Gabler, Wiesbaden.

Malpezzi, S. (2003), "Hedonic pricing models: a selective and applied review", *Housing Economics and Public Policy*, Vol. 1, pp. 67–89.

Mason, C. and Quigley, J.M. (1996), "Non-parametric hedonic housing prices", *Housing Studies*, Vol. 11 No. 3, pp. 373–385.

Matysiak, G. (2017), "Automated Valuation Models (AVMs): A brave new world?", paper presented at Wroclaw Conference in Finance 2017, available at: https://www.researchgate.net/profile/george-matysiak/publication/319355261_automated_valuation_models_avms_a_brave_new_world.

Mayer, M., Bourassa, S.C., Hoesli, M. and Scognamiglio, D. (2019), "Estimation and updating methods for hedonic valuation", *Journal of European Real Estate Research*, Vol. 12 No. 1, pp. 134–150.

McCluskey, W. and Anand, S. (1999), "The application of intelligent hybrid techniques for the mass appraisal of residential properties", *Journal of Property Investment & Finance*.

McCluskey, W.J., McCord, M., Davis, P.T., Haran, M. and McIlhatton, D. (2013), "Prediction accuracy in mass appraisal: a comparison of modern approaches", *Journal of Property Research*, Vol. 30 No. 4, pp. 239–265.

Metzner, S. and Kindt, A. (2018), "Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties", *International Journal of Housing Markets and Analysis*, Vol. 11 No. 1, pp. 73–100.

Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the Analysis of Survey Data, and a Proposal", *Journal of the American statistical association*, 58(302), 415-434.

Nelder, J.A. and Wedderburn, R.W.M. (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society. Series A (General)*, Vol. 135 No. 3, p. 370.

Nobis, C. and Kuhnimhof, T. (2018), Mobilität in Deutschland − MiD: Ergebnisbericht.

Oust, A., Hansen, S.N. and Pettrem, T.R. (2020), "Combining Property Price Predictions from Repeat Sales and Spatially Enhanced Hedonic Regressions", *The Journal of Real Estate Finance and Economics*, Vol. 61 No. 2, pp. 183–207.

Pace, K. (1998), "Appraisal Using Generalized Additive Models", *Journal of Real Estate Research*, Vol. 15 No. 1, pp. 77–99.

Pace, R.K. and Hayunga, D. (2020), "Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests", *The Journal of Real Estate Finance and Economics*, Vol. 60 No. 1-2, pp. 170–180.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T. and French, N. (2003), "Real estate appraisal: a review of valuation methods", *Journal of Property Investment & Finance*, Vol. 21 No. 4, pp. 383–401.

Peterson, S. and Flanagan, A. (2009), "Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal", *Journal of Real Estate Research*, Vol. 31 No. 2, pp. 147–164.

Powe, N.A., Garrod, G.D. and Willis, K.G. (1995), "Valuation of urban amenities using an hedonic price model", *Journal of Property Research*, Vol. 12 No. 2, pp. 137–147.

Prajwala, T.R. (2015), "A Comparative Study on Decision Tree and Random Forest Using R Tool", IJARCCE, pp. 196–199.

Quinlan, J.R. (1979), "Discovering rules by induction from large collections of examples", In D. Michie (Ed.), Expert systems in the micro electronic age. Edinburgh University Press.

Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy*, Vol. 82 No. 1, pp. 34–55.

Rossini, P. and Kershaw, P. (2008), "Automated valuation model accuracy: Some empirical testing", (Doctoral dissertation, Pacific Rim Real Estate Society).

Sagi, O. and Rokach, L. (2018), "Ensemble learning: A survey", WIREs Data Mining and Knowledge Discovery, Vol. 8 No. 4.

Sangani, D., Erickson, K. and Hasan, M.A. (2017), "Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting", in 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), IEEE.

Schulz, R. and Wersing, M. (2021), "Automated Valuation Services: A case study for Aberdeen in Scotland", *Journal of Property Research*, Vol. 38 No. 2, pp. 154–172.

Schulz, R., Wersing, M. and Werwatz, A. (2014), "Automated valuation modelling: a specification exercise", *Journal of Property Research*, Vol. 31 No. 2, pp. 131–153.

Singh, A., Sharma, A. and Dubey, G. (2020), "Big data analytics predicting real estate prices", International *Journal of System Assurance Engineering and Management*, Vol. 11 No. S2, pp. 208–219.

Sirmans, G.S., Macpherson, D.A. and Zietz, E.N. (2005), "The Composition of Hedonic Pricing Models", *Journal of Real Estate Literature*, Vol. 13 No. 1, pp. 1–44.

Tchuente, D. and Nyawa, S. (2021), "Real estate price estimation in French cities using geocoding and machine learning", *Annals of Operations Research*, pp. 1–38.

Thompson, J.F. and Gordon, J.F. (1987), "Constrained regression modeling and the multiple regression analysis-comparable sales approach", *Property Tax Journal*, 6(4), 251-262.

Tianqi, C. and Guestrin, C. (2016), XGBoost: A scalable tree boosting system.

Todora, J. and Whiterell, D. (2002), “Automating the sales comparison approach”, *Assessment Journal*, 9(1), 25-33.

Trawinski, B., Telec, Z., Krasnoborski, J., Piwowarczyk, M., Talaga, M., Lasota, T. and Sawilow, E. (2017), “Comparison of expert algorithms with machine learning models for real estate appraisal”, in 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), IEEE.

Underwood, W.E. and Moesch, J.R. (1982), “The second generation of CAMA in New York State”, In First World Congress on Computer Assisted Valuation, Cambridge, MA.

Wood, S.N. (2017), Generalized Additive Models: An Introduction with R, Second Edition, CRC Press.

Yang, J., Bao, Y., Zhang, Y., Li, X. and Ge, Q. (2018), “Impact of Accessibility on Housing Prices in Dalian City of China Based on a Geographically Weighted Regression Model”, *Chinese Geographical Science*, Vol. 28 No. 3, pp. 505–515.

Zurada, J., Levitan, A. and Guan, J. (2011), “A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context”, *Journal of Real Estate Research*, Vol. 33 No. 3, pp. 349–388.

4 Explainable AI in a Real Estate Context – Exploring the Determinants of Residential Real Estate Values

4.1 Abstract

A sound understanding of real estate markets is of economic importance and not simple, as properties are a heterogenous asset and no two are alike. Traditionally, parametric or semi-parametric and, thus, assumption-based hedonic pricing models are used to analyze real estate market fundamentals. These models are characterized by the fact that they require a-priori assumptions regarding their functional form. Usually, the true functional form is unknown and characterized by non-linearities and joint effects, which are hard to fully capture. Therefore, the validity of their results is limited. Applying the state-of-the art non-parametric machine learning XGBoost algorithm, in combination with the model-agnostic Accumulated Local Effects Plots, (ALE) enables us to overcome this problem. Using a dataset of 81,166 residential properties for the seven largest German cities, we show how this enables us to analyze the value-determining effects of several structural, locational and socio-economic hedonic features. Our findings lead to a deeper representation of real estate market fundamentals.

Keywords: Housing Market, Machine Learning, Explainable AI, Feature Importance, ALE Plots

4.2 Introduction

Understanding real estate markets and its drivers is arguably one of the most important areas of real estate research. Compared to other asset classes, real estate is a heterogeneous asset that differs from one another in terms of its features. There is a large body of literature dealing with the factors which have a significant influence on the value or price of a property, subsumed under the term Hedonic Price Models. They are usually based on parametric and semi-parametric methods like the Ordinary Least Square approach (see, e.g., Malpezzi (2003), Sirmans et al. (2005) and Schulz et al. (2014)) or the Generalized Additive Models (see, e.g., Mason and Quigley (1996), K. Pace (1998), Bao and Wan (2004), Bourassa et al. (2007), Bourassa et al. (2010) and Brunauer et al. (2010)).

In recent years, pushed by increasing digitalization, data availability and growing computing capacity, new methods have entered the scene. So-called modern Machine Learning (ML) methods are now being used in a wide variety of areas and are also finding their way into the real estate sector. Several studies have already focused on the application of ML in the context of real estate valuation. The results show that ML algorithms provide a higher level of valuation accuracy than well-established parametric and semi-parametric methods (see, e.g., Chun Lin and Mohan (2011), Kok et al. (2017) and Mayer et al. (2019)). This indicates that real estate markets are characterized by non-linearity and joint effects, which ML methods can efficiently capture and exploit. Furthermore, ML algorithms are often less restrictive in terms of their model structure and thus are more flexible.

However, ML applications are usually considered as so-called black boxes. While parametric and semi-parametric applications are comprehensible to humans, the calculations of modern ML applications can only be understood with difficulty if at all. This is where the research on eXplainable Artificial Intelligence (XAI) comes in: using model-agnostic approaches, the modes of operation of ML algorithms are revealed and thus become transparent in their mode of action.

In real estate, XAI approaches have been explored only to a limited extend, but can yield several benefits. First, they might support research in understanding the key drivers of real estate markets by taking non-linearity and joint effects into consideration. There is great potential in the in-depth analysis of the mechanisms between real estate market values or prices and their fundamentals. On the one hand, the dependency between property values and their value-determining features such as amenities and points of interests can be analyzed in detail. On the other hand, the relationships between these features can also be investigated. These analyses might provide deeper insights into (local) real estate market drivers, which are relevant for different market actors, for example investors, lenders, and real estate developer. Second, XAI methods shed light on the mechanism of ML algorithms, thus overcoming their image of black boxes, and thus increasing their acceptance in regulated areas, for example in the mortgage lending industry.

In this study, we rely on two different XAI techniques to render the deep hidden patterns of residential real estate market values interpretable to human beings, namely Permutation Feature Importance (PFI), first introduced by Breiman (2001), and Accumulated Local Effects Plots (ALE), established by Apley and Zhu (2020). PFI is used to analyze which features actually influence the value of a property. ALE plots allow us to make statements about the effects themselves and whether non-linear relationships can be identified or not. In particular, the former is used as a basis for the latter, to identify which variables have the greatest impact on property values.

In XAI research, Partial Dependence Plots (PDP) – implemented by Friedman (2001) – are one of the oldest and most widely used methods. However, PDP plots have been shown to produce biased results when features are dependent (Apley and Zhu (2020)). In real estate, many features have an intrinsic dependence that does not justify the use of PDP plots. In contrast, ALE plots do not have this disadvantage and are therefore well suited to real estate market analysis. To the best of our knowledge, we are the first to use ALE plots in context of real estate market analysis and thus, to provide unbiased evidence of underlying factors and their relationship.

Based on the ML algorithm eXtreme Gradient Boosting (XGBoost), a unique dataset of 81,166 residential properties for the Top-7 cities of Germany is used for our analysis. The dataset can be split into 61,763 condominiums and 19,403 single-family homes. We analyze the two groups separately in order to reveal differences between the two property subtypes, in addition to the general analysis of the value-determining features. We use the market values of the single properties as our target variable. These market values are based on appraiser valuations and are therefore objectively verified by professional real estate evaluators.

Besides the general introduction of ALE plots in a real estate context, we contribute to the literature by addressing the following research questions:

I. Which characteristics are important for the market values of residential properties?
II. To what extent are the features characterized by either linearity or non-linearity? Are there differences depending on different cities?
III. Are there fundamental differences between condominiums and single-family homes?

Our analyses reveal that the same value-determining features play a predominant role for both condominiums and single-family homes. However, the actual effects of each feature are often different for the two property subtypes. Furthermore, we identify non-linear relationships for the majority of features. Generalized rules of thumb such as "the larger the living area, the lower the market value per square meter" are refuted by our findings for condominiums, but can be confirmed for single-family homes. In summary, our results show how important it is for both real estate research and practice to conduct data-driven analyses with the help of modern ML and XAI approaches, in order to gain important market insights and, if necessary, to update long-established assumptions regarding the determinants of real estate market values.

4.3 Literature Review

Hedonic Price Models are widely used in real estate research. A hedonic price function explains the price or value variations of different properties on the market, by the differences in their feature characteristics. As described by Malpezzi (2003), the most common one is the Ordinary Least Squared regression (OLS). However, this approach is associated with some limitations due to the underlying assumptions of linearity, the lack of flexibility, multicollinearity, and spatial autocorrelation (Zurada et al. (2011)). The latter refers to the clustering of similar real estate values within a geographical space or submarket, as described by Can (1990). Several spatial econometric models have been developed over recent decades (see, e.g,, Bourassa et al. (2010)). Most show that hedonic models correcting for spatial autocorrelation deliver better results than an OLS. For example the spatial general model (SAC), the spatial autoregressive model (SAR), the spatial error model (SEM) and the spatial Durbin model (SDM), (see, e.g., Dubin (1988), Conway et al. (2010) and Stamou et al. (2017)).

Not only the discussion of which method to use, but also which features to analyze is actively debated in the real estate literature. These include ecological circumstances such as air pollution (Fernández-Avilés et al. (2012), Simons et al. (2015) and Lu and Lee (2022)), water restrictions (Carstens et al. (2020)) and land erosion (Below et al. (2015) and Dumm et al. (2018)). The effect of different amenities is also often the subject of various studies, for example, the proximity of wind turbines (Hoen and Atkinson-Palombo (2016)), public schools (Des Rosiers et al. (2001) and Hwang et al. (2019)), public and private green spaces (Turner and Seo (2021)), power lines (Wyman and Mothorpe (2018)) or waterfront properties (Dumm et al. (2016) and Jauregui et al. (2019)).

In recent years, more advanced statistical and machine learning methods have gained interest in the real estate community, as they are able to overcome almost most issues of the OLS. Especially deep learning algorithms like Artificial Neural Networks (ANN), bagging techniques like random forest (RF) and boosting algorithms like the eXtreme Gradient Boosting (XGBoost) algorithm, seem to be better suited to real-estate-related problems. Applications include Worzala et al. (1995), Din et al. (2001), Peterson and Flanagan (2009), McCluskey et al. (2013) and Chiarazzo et al. (2014) for neural networks; Antipov and Pokryshevskaya (2012), Bogin and Shui (2020) and R. K. Pace and Hayunga (2020) for random forests. Focusing on boosting related methods, see van Wezel et al. (2005), Kagie and van Wezel (2007), Gu and Xu (2017), Sangani et al. (2017), Ho et al. (2021) and Stang et al. (2021). In almost all cases, researchers conclude that machine learning techniques yield better predictions than standard linear models. However, these methods have been criticized for their lack of transparency (Din et al. (2001) and McCluskey et al. (2013)).

This criticism paves the way for a new stand of literature which focuses on the explainability of machine learning in real estate. In many studies, Partial Dependence Plots (PDP), following Friedman (2001), are used, see, for example, Mayer et al. (2019) and Lorenz et al. (2021). PDP plots are calculated by varying each feature over its marginal distribution, i.e., observed values, holding all other features constant and re-predicting the target variable. This approach is reliable if all features are independent, otherwise the PDP plots are severely biased. However, in real estate applications, many features are inherently dependent. For example, the living area and the number of rooms are intrinsically interdependent. In the calculation of the PDP plots, one would incorporate unrealistic data pairs, such as a house with 1 room and

400 square meters or a house with 10 rooms and 40 square meters. An excellent discussion on this issue can be found in Apley and Zhu (2020).

Therefore, this paper extends the literature by using a novel method – namely the ALE plots - to reveal feature relations and deal with the specialties of real estate research. Following Apley and Zhu (2020), ALE plots investigate the way features affect, on average, the prediction of an ML model and can be seen as an unbiased substitute for PDPs.

Furthermore, the literature has mainly focused on the comparison in terms of accuracy, but falls short on the economic implications of non-linearity in real estate research. Accordingly, this paper offers reliable and unbiased relations between features and house prices and discusses their economic implications. Furthermore, we disentangle these relationships on regional basis in order to evaluate structural differences. This is especially important for mortgage underwriters, valuation firms and regulatory authorities and, thus, of considerable interest to most of the real estate community.

4.4 Data

For the purpose of this study, a dataset consisting of 81,166 residential properties for the Top-7 cities of Germany is used. The Top-7 are the most important cities in Germany for the real estate industry and are: Berlin, Dusseldorf, Frankfurt am Main, Hamburg, Cologne, Munich and Stuttgart. In comparison to other European countries such as England or France, the German real estate market is polycentric and not dominated by one large city. As Cajias and Freudenreich (2018) explain, analyzing the Top-7 cities leads to a "socially, culturally and economically well diversified overview of major urban areas all over Germany". As we are interested in analyzing differences between different subtypes of residential properties, the dataset is further split into two groups. The first consists of 61,763 condominiums and the second of 19,403 single-family homes. Table 4.1 shows how the individual observations are distributed among the seven cities.

Table 4.1: Observations per city and subtype

	Berlin	Dusseldorf	Frankfurt	Hamburg	Cologne	Munich	Stuttgart
Condominiums	15,166	5,295	5,559	5,703	13,189	12,743	4,108
Single Family Homes	6,545	814	1,140	3,555	4,933	1,408	1,008

The dataset is provided by a large German banking group and originates from their valuation department. The data was collected mainly for lending purposes between 2014 and 2020. The market value per square meter of the properties, based on professional appraiser valuations, is used as the target variable. In contrast to listing data, market values do not depend on subjective seller perceptions of value, but are assessed objectively by outside third parties. In addition to the dependent variable, a set of features defining the structural characteristics of the properties is used. All properties are georeferenced, making it possible to add a spatial gravity layer to account for spatial information. Features describing the location and neighborhood of the observations are added via Open Street Map and Acxiom[8]. The dataset is cleaned before being used to account for duplicates, incompleteness and erroneous data points. Table 4.2 summarizes the descriptive statistics of the features used for condominiums and Table 4.3 those for single-family homes.[9]

[8] Acxiom is an American data provider for international data. Further information can be found at: https://www.acxiom.com/.

[9] The individual summary statistics for each city are available on request.

Table 4.2: Condominium – Descriptive statistics

Variable	Unit	Mean	Median	Standard Deviation	Maximum	Minimum
Market value per square meter	Float	3,691.20	3,254.55	1,911.97	1,8384.40	216.96
Living area	Float	72.19	69.00	28.34	203.57	15.0
Longitude	Float	10.06	10.00	2.52	13.73	6.70
Latitude	Float	50.80	50.95	1.79	53.71	48.07
Micro score - education and work	Float	94.59	97.88	7.58	99.89	0.00
Micro score - shopping	Float	88.84	92.78	11.04	99.29	0.00
Micro score - leisure	Float	98.84	99.64	3.35	99.98	0.00
Micro score - public transport	Float	64.22	67.86	19.22	97.90	0.00
Year of construction	Integer	1974	1973	33.80	2023	1900
Year of valuation	Integer	2016	2017	2.02	2020	2014
Quarter of valuation	Integer	2.45	2.00	1.12	4.00	1.00
Quality grade	Integer	3.19	3.00	0.52	5.0	1.00
Time on market	Integer	10.50	9.90	4.23	60.7	2.80
Unemployment ratio	Integer	6.27	5.60	4.42	26.89	0.04
Condition very good	Binary	0.18	0.00	0.38	1.00	0.00
Condition good	Binary	0.39	0.00	0.48	1.00	0.00
Condition middle	Binary	0.45	0.00	0.50	1.00	0.00
Condition moderate	Binary	0.01	0.00	0.11	1.00	0.00
Condition bad	Binary	0.00	0.00	0.04	1.00	0.00
Owner-occupied & Non-owner-occupied	Binary	0.10	0.00	0.31	1.00	0.00
Owner-occupied	Binary	0.44	0.00	0.50	1.00	0.00
Non-owner-occupied	Binary	0.46	0.00	0.50	1.00	0.00

Table 4.3: Single-family homes – Descriptive statistics

Variable	Unit	Mean	Median	Standard Deviation	Maximum	Minimum
Market value per square meter	Float	3,064.06	2,693.19	1,538.35	2,2781.21	199.44
Living area	Float	133.68	126.43	42.19	402.00	30.77
Lot size	Float	467.92	396.00	296.63	3500.00	1.00
Longitude	Float	10.25	10.03	2.64	13.75	6.70
Latitude	Float	51.60	52.40	1.59	53.714	47.58
Micro score - education and work	Float	85.59	88.29	11.87	99.83	0.00
Micro score - shopping	Float	75.20	79.49	15.99	98.88	0.00
Micro score - leisure	Float	95.49	98.27	8.75	99.98	0.00
Micro score - public transport	Float	43.28	42.78	16.48	95.37	0.00
Year of construction	Integer	1974	1977	30.18	2022	1900
Year of valuation	Integer	2016	2016	1.97	2020	2014
Quarter of valuation	Integer	2.44	2.00	1.11	4.00	1.00
Quality grade	Integer	3.15	3.00	0.50	5.00	1.00
Time on market	Integer	11.30	10.20	3.71	60.70	3.70
Unemployment ratio	Integer	8.34	9.44	4.33	26.89	0.08
Condition very good	Binary	0.15	0.00	0.36	1.00	0.00
Condition good	Binary	0.42	0.00	0.49	1.00	0.00
Condition middle	Binary	0.41	0.00	0.49	1.00	0.00
Condition moderate	Binary	0.02	0.00	0.13	1.00	0.00
Condition bad	Binary	0.00	0.00	0.04	1.00	0.00
Basement	Binary	0.19	0.00	0.39	1.00	0.00
No basement	Binary	0.81	1.00	0.39	1.00	0.00
Owner-occupied & Non-owner-occupied	Binary	0.17	0.00	0.37	1.00	0.00
Owner-occupied	Binary	0.74	1.00	0.44	1.00	0.00
Non-owner-occupied	Binary	0.09	0.00	0.28	1.00	0.00
Detached house	Binary	0.41	0.00	0.49	1.00	0.00
Non-detached house	Binary	0.59	1.00	0.49	1.00	0.00

In the area of structural characteristics, the construction year, living area, use of the property, condition and a quality grade were used for both apartments and single-family homes. Furthermore, the lot size, a variable describing whether the property has a basement or not and a feature outlining the subtype of the property, are used for the single-family homes. All these features were determined by professional appraisers in the context of their assessment process, which is why it can reasonably be assumed that these represent a detailed and truthful representation of the actual properties.

The use of the building describes the possible types of usage of the property, whereby the characteristics are either “Owner-occupied & Non-owner-occupied”, “Owner-Occupied” or

"Non-owner-occupied". Basically, the variable describes whether or not a property can be rented to a third-party. The quality of the property is measured via a grade, on a scale ranging from 1 (very poor) to 5 (very good). The general condition of the property is represented by a categorial variable with 5 different categories ranging from bad to very good. The features describing the subtype of the single-family homes are binary, and state whether it's a detached or non-detached house.

Features representing the micro-location of a property are the latitude and longitude and the four different micro scores. The micro scores of a location are calculated via a gravity model and reflect the accessibility as the proximity to selected everyday destinations for each category. A gravity model is a common method for approximating the accessibility of a location and is based on the assumption that nearby destinations play a greater role in everyday life than more distant ones (Handy and Clifton (2001)). The scores can range from 0 to 100 points, and the higher the score, the better the accessibility of the location. The relevant points-of-interest (POIs) are selected on the basis of the findings of Powe et al. (1995), Metzner and Kindt (2018), Yang et al. (2018), Nobis and Kuhnimhof (2018) and Huang and Dall'erba (2021) and are provided in Table 4.4. The scores are mainly used to reduce dimensionality and complexity. A more detailed description of the construction of the micro scores can be found in Appendix I.

Table 4.4: Features of the micro scores of a location

Points-of-Interests	Category	Description
University	Education & Work	University campus: an institute of higher education
School	Education & Work	Place for education
Kindergarten	Education & Work	Facility for early childhood care
CBD	Education & Work	Center of the next city
Supermarket	Local Supply	Supermarket – a large store with groceries
Marketplace	Local Supply	A marketplace where goods are traded daily or weekly
Chemist	Local Supply	Shop focused on selling articles of personal hygiene, cosmetics, and household cleaning products
Bakery	Local Supply	Place for fresh bakery goods
ATM	Local Supply	ATM or cash point
Hospital	Local Supply	Facility providing in-patient medical treatment
Doctors	Local Supply	Doctor's practice / surgery
Pharmacy	Local Supply	Shop where a pharmacist sells medications
Restaurant	Leisure & Food	Facility to go out to eat
Café	Leisure & Food	Place that offers casual meals and beverages
Park	Leisure & Food	A park, usually urban (municipal)
Fitness Centre	Leisure & Food	Fitness Centre, health club or gym
Movie Theater	Leisure & Food	Place where films are shown
Theater	Leisure & Food	Theatre or opera house where live performances occur
Shopping Mall	Leisure & Food	Shopping Centre– multiple stores under one roof
Department Store	Leisure & Food	Single large store selling a large variety of goods
Subway Station	Transportation	City passenger rail service
Tram Station	Transportation	City passenger rail service
Railway Station	Transportation	Railway passenger-only station.
Bus Stop	Transportation	Bus stops of the local bus lines.
E-Charging Station	Transportation	Charging facility for electric vehicles

Note: The descriptions of the selected Points-of-Interests is based on the explanations of Open Street Map.[10]

The macro-location is considered by means of the features “Unemployment Rate” and “Time-On-Market”. These two features have been used frequently in other studies (see, e.g., Cheng and Fung (2015)). Both variables are available at ZIP Code level. The “Unemployment Rate” measures the percentage of workers in the labor force who do not currently have a job, but are actively looking for work, and is used as a proxy for the social status of the local inhabitants. The feature “Time-On-Market” is used as a proxy for liquidity and is defined as the average number of days properties are advertised on the market within a certain ZIP Code.

[10] See https://wiki.openstreetmap.org/wiki/Map_features.

To capture a temporal trend and seasonality, the year and quarter of the valuation are included. There are no correlations of concern within the data set, so that all variables can be integrated accordingly.[11]

4.5 Methodology

Extreme Gradient Boosting – XGBoost

Since it yielded reasonable results in several research articles (see, e.g., Truong et al. (2020)), the XGBoost is chosen as our underlying ML model. Especially Stang et al. (2021) showed that the XGBoost archived the best results for estimating real estate market values in Germany. Therefore, the XGBoost ensures a good model-fit and enables a post-hoc analysis of the results and the application of the PFI and the ALE plots. The XGBoost is a tree-based ensemble learning method. These algorithms combine many so-called weak learners, h_m, in our case, single decision trees, into one strong learner h:

$$h(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} u_m h_m(\boldsymbol{y}|\boldsymbol{x})\,.$$

where u_m is used to weight the weak learners. M is the number of single trees, $\boldsymbol{x}$ is the full features space and $\boldsymbol{y}$ the response variable. Boosting is a type of ensemble learning in which the weak learners h_m are trained sequentially. Starting with one tree, the following models learn from the previous errors. Gradient boosting uses the so-called gradient decent algorithm by adding new trees to minimize the loss of the model. The eXtreme Gradient Boosting is a computationally effective and highly efficient version of gradient boosting. The advantage of XGBoost is that it can recognize very complex patterns within large amount of data. For more information about tree-based methods, ensemble learning and gradient boosting, the interested reader is recommended to Hastie et al. (2001).

[11] The correlation matrix is available on request.

Testing concept

In order to evaluate the XGBoost, five-fold cross validation is used. To obtain the overall performance, we use the set of evaluation metrics presented in Table 4.5. The selected metrics are applied continuously, to evaluate the results of hedonic and machine learning approaches (see, e.g., Mayer et al. (2019)).

Table 4.5: Evaluation metrics

Error	Formula	Description
Mean Absolute Percentage Error (MAPE)	$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	Mean of all absolute percentage errors. A lower MAPE signals higher prediction accuracy in percent.
Median Absolute Percentage Error (MdAPE)	$MdAPE(y, \hat{y}) = median(\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert)$	Median of all absolute percentage errors. A lower MdAPE denotes a higher precision in percent without being sensitive to outliers.
Error buckets (PE(x))	$PE(x) = 100\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert < x$	Percentage of predictions where the relative deviation is less than x%, with x being 10 and 20. A larger PE(x) signals a lower variation in the predictions.
R^2	$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Coefficient of determination. A high R^2 is an indication of better goodness of fit of the model.

Permutation Feature Importance

Permutation feature importance (PFI) is a post-hoc global model-agnostic technique for detecting the influence of the features used on the predictions. We use the PFI to first identify the most important features before analyzing them in more detail afterwards. One main advantage of the PFI is that it can be applied to all machine learning models. Initially introduced only for random forest by Breiman (2001), Fisher et al. (2019) developed this method further to be applicable to all models and called it 'model reliance'. A feature is considered as important if the prediction error increases after its entries are permuted. Therefore, the permutation feature importance of feature j can be defined as:

$$PFI_j = E\left(L\big(\hat{f}(\boldsymbol{x}_j, \boldsymbol{x}_{-j}), \boldsymbol{y}\big)\right) - E(L\left(\hat{f}(\boldsymbol{x}, \boldsymbol{y})\right),$$

where L denotes a chosen loss function, f refers to a fitted supervised machine learning model, $\boldsymbol{x}_j$ and $\boldsymbol{x}_{-j}$ are the permuted variable j and its complementary set of features. Furthermore, $\boldsymbol{x}$ defines the full features space and $\boldsymbol{y}$ the response variable. In this paper, the Mean Absolute Percentage Error is used as a loss function. For every feature j, the permutation feature importance is computed 100 times, each time randomly permuting its entries. To obtain the final PFI of j, hundred permutation feature importances are averaged.

Accumulated local effects

Accumulated local effects (ALE) plot, developed by Apley and Zhu (2020), is a feature effect approach that is able to analyze how a single feature influences predictions for any ML model. Therefore, the averages of changes in the predicted values dependent on the conditional distribution of the single feature, is computed. Assume that $\boldsymbol{x}$ is the full feature space containing d variables and $\boldsymbol{y}$ the response variable. $\hat{f}$ is a fitted supervised machine learning model that is differentiable and uses $\boldsymbol{x}$ to predict $\boldsymbol{y}$. Define $\boldsymbol{x}_j$ as the feature of interest and $\boldsymbol{x}_{-j}$ the complementary set of features, $j \in \{1, \dots d\}$. X_j represents the j^{th} feature as a random variable. Then, the ALE main effect of $\boldsymbol{x}_j$ can be calculated:

$$\hat{f}_{j,ALE} = \int_{z_{0,j}}^{x_j} E\left[\frac{\partial \hat{f}(X_j, X_{-j})}{\partial X_j} \mid X_j = z_j\right] dz_j - constant,$$

with $z_{0,j}$ being a lower bound of X_j. Usually, $z_{0,j}$ is defined as $\min\{\boldsymbol{x}_j\}$. The expected value E is computed conditional on the representation of $\boldsymbol{x}_j$ and over the marginal distribution of $\boldsymbol{x}_{-j}$. The constant is subtracted to center the plot.

Since not every machine learning model is differentiable, Apley and Zhu (2020) introduced a way to estimate the ALE for any supervised machine learning model. Therefore, the value range of the j^{th} feature is divided into K intervals $N_j(k)$, $\{N_j(k) = \left(z_{k-1,j}, z_{k,j}\right]: k = 1,2, \dots K\}$, where $z_{j,k}$ refers to the upper and $z_{k-1,j}$ the lower boundary of interval k.

Furthermore, x^* is a specific value of $\boldsymbol{x}_j$ and $k_j(x^*)$ denotes the index of the interval x^* belongs to. $n_j(k)$ is the number of observations in each interval k and $\boldsymbol{x}_{i,-j}$ represents the observations of the remaining features, $i \in \{1, 2, \dots N\}$.

Before we can compute the main effect $\hat{f}_{j,ALE}$, the uncentered ALE $\hat{g}_{j,ALE}$ of the j^{th} feature has to be calculated for every $x^* \in (z_{0,j}, z_{K,j}]$, where $z_{0,j}$ is just below the minimum observation of $\{x_{i,j}: i = 1, \dots n\}$ and $z_{K,j}$ is the maximum observation of $\{x_{i,j}: i = 1, \dots n\}$:

$$\hat{g}_{j,ALE}(x^*) = \sum_{k=1}^{k_j(x^*)} \frac{1}{n_j(k)} \sum_{\{i: x_{i,j} \in N_j(k)\}} \left[\hat{f}\left(z_{k,j}, \boldsymbol{x}_{i,-j}\right) - \hat{f}\left(z_{k-1,j}, \boldsymbol{x}_{i,-j}\right)\right].$$

The ALE main effect estimator can be now computed by subtracting an estimate of $E[g_{j,ALE}(X_j)]$:

$$\hat{f}_{j,ALE}(x^*) = \hat{g}_{j,ALE}(x^*) - \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{j,ALE}\left(x_{i,j}\right) = \hat{g}_{j,ALE}(x^*) - \frac{1}{n}\sum_{k=1}^{K} n_j(k) \cdot \hat{g}_{j,ALE}\left(z_{k,j}\right).$$

The ALE plots have many advantages. Among other things, they are fast to compute and unbiased. Therefore, they can be used even if features are correlated, in contrast to partial dependency plots. Besides that, the ALE plots are centered so that the mean effect of the features is zero. Therefore, the y-Axis of the ALE can be interpreted as the main effect of the independent variable at a certain point, in comparison to the average predicted value. For further information about the ALE plot, we recommend having a look at Apley and Zhu (2020).

4.6 Results

An extra XGBoost model was trained for each of the seven cities. Furthermore, different algorithms were trained for the condominiums and the single-family homes. We apply random cross validation with 5 folds for the model validation. Table 4.6 shows the average results of the evaluation metrics across all cities. The results for each city can be seen in Appendix II.

Table 4.6: Results XGBoost for all Top-7 cities

	XGBoost		OLS	
Metrics	Single-family homes	Condominiums	Single-family homes	Condominiums
MAPE	0.1441	0.1253	0.1834	0.1986
MdAPE	0.0988	0.0829	0.1314	0.1467
PE(10)	0.5073	0.5698	0.3999	0.3654
PE(20)	0.7861	0.8247	0.6833	0.6396
R^2	0.7133	0.8117	0.5623	0.6230

To ensure that we can make a statement about the quality of the results, we estimated a basic OLS regression for each city and can thus benchmark the results. Our results indicate strong and robust model performance across all Top-7 cities for the XGBoost. For all metrics, the XGBoost yields a better result than the OLS, which could be expected from the literature. The MAPE of the XGBoost, relative to the OLS, is 37% lower for condominiums and 21% lower for single-family homes. This difference can already serve as a first indicator of non-linearities, joint effects and higher order interactions within the data. The trained XGBoost algorithms are therefore well suited to a post-hoc analysis of the results and the application of the PFI and the ALE plots.

Results Permutation Feature Importance (PFI)

In a first step, we use the PFI to determine which variables are important for predicting the market value per square meter. The PFI provides a highly compressed, global insight into the machine learning model's behavior. The PFI is easy to interpret and also takes into account interactions within the individual features, as described by Molnar (2020). The PFI ranks all features used in the model according to their influence on the dependent variable. Therefore, in our case, the higher the ranking of a feature, the greater its influence on the market value per square meter of a property. Table 4.7 shows the five highest ranked features for condominiums in each of the Top-7 cities. Table 4.8 shows this for single-family homes.

Table 4.7: Top-5 features per city – Condominiums

	Berlin	Dusseldorf	Frankfurt	Hamburg	Cologne	Munich	Stuttgart
Top-1 Feature	Year of valuation	Year of construction	Year of construction	Year of valuation	Year of valuation	Year of valuation	Year of valuation
Top-2 Feature	Year of construction	Year of Valuation	Year of valuation	Unemploy-ment ratio	Year of constructio n	Year of constructio n	Year of constructio n
Top-3 Feature	Unemploy-ment ratio	Longitude	Longitude	Year of construction	Unemploy-ment ratio	Longitude	Unemploy-ment ratio
Top-4 Feature	Longitude	Latitude	Unemploy-ment ratio	Longitude	Longitude	Living area	Living area
Top-5 Feature	Latitude	Living area	Latitude	Latitude	Latitude	Unemploy-ment ratio	Longitude

Table 4.8: Top-5 features per city – Single-family homes

	Berlin	Dusseldorf	Frankfurt	Hamburg	Cologne	Munich	Stuttgart
Top-1 Feature	Year of valuation	Lot size	Year of valuation	Year of valuation	Year of valuation	Year of valuation	Year of valuation
Top-2 Feature	Year of construction	Year of valuation	Lot size	Lot size	Living area	Lot size	Lot size
Top-3 Feature	Lot size	Year of construction	Living area	Unemploy-ment ratio	Lot size	Living area	Living area
Top-4 Feature	Living area	Living area	Longitude	Living area	Unemploy-ment ratio	Year of constructio n	Year of constructio n
Top-5 Feature	Longitude	Unemploy-ment ratio	Year of construction	Year of construction	Year of constructio n	Latitude	Unemploy-ment ratio

The PFI analysis shows that the same features play a predominant role for both condominiums and single-family homes. The valuation year is by far the most important. For condominiums, it is always at the top of the list except in Dusseldorf and Frankfurt, and for single-family homes it is also always the highest ranked feature except in Dusseldorf. This finding is interesting, in that the year of valuation feature is purely a factor describing the market phase. It shows that the market values of the properties are mainly influenced by the market phase and thus the general market trends. Their structural characteristics are initially not in the foreground. In addition to the valuation year, it can be seen that structural, location-related and socio-economic features have an important influence on the market value of the properties. These findings are also in line with Dubin (1988) and Sirmans et al. (2005). Not only structural features, but also the location and the economic or social environment of a property are decisive for the composition of market values in hedonic pricing models. In the case of condominiums, the year of construction and the living area are the most important factors in terms of property characteristics. In the case of single-family homes, lot size is added to these features. This is also in line with the general findings of other studies (see, e.g., Fan et al. (2006)). In our case, the location of the properties is represented by the latitude and longitude. It turns out that depending on the location, the values for otherwise identical properties are different. Socio-economically, we find that the unemployment ratio seems to play a partially important role.

Results Accumulated Local Effects Plots (ALE)

To analyze the identified features in more detail, we use ALE plots to take a closer look at how the individual effects work and what economic insights can therefore be drawn. ALE plots describe the main effect of a feature at a certain point in comparison to the average predicted value. Compared to traditional hedonic price functions, they can capture non-linearities independently, without the need for a priori manual specification. This enables us to visualize a more realistic representation of the actual market fundamentals. Our findings are therefore beneficial for all real estate market actors, to reach more targeted and, in particular, data-supported decisions.

The ALE plot is centered and the mean effect of the features is zero. Therefore, the y-axis of the ALE plot can be interpreted as the main effect of the independent variable at a certain point, in comparison to the average predicted value. The ALE algorithm divides the feature space into intervals containing the same number of data points, whereby feature intervals with a greater observation density are chosen to be smaller than intervals with a low density. In our case, the maximum number of intervals is set to 250. In contrast to other model-agnostic approaches, the ALE can handle correlated features and therefore deliver robust and unbiased results. In order to check whether there are non-linearities within the data, we show within the ALE plots, in addition to the effect identified by the XGBoost, the results of a basic OLS as a benchmark. The results of the OLS are shown by means of a red line and show the main difference between previous hedonic pricing approaches and to the results of our analysis. We are thus able to show which effects are covered well by parametric models and which effects can be covered rather poorly.

Starting with the structural features, Figure 1 shows the ALE plots for the year of construction for condominiums, and Figure 2 for single-family homes. In contrast to the red OLS line, it is obvious that the trend is not linear. The effects are approximately the same across all cities. Comparing the graphs of condominiums with those of single-family homes, we see that the effects are essentially the same. It is notable that the negative trend for middle-aged properties is more pronounced for the former, suggesting that the year of construction has a generally greater influence on condominiums. Traditionally, the effect of the year of construction is described and incorporated as u-shaped (see, e.g., Mayer et al. (2019)), as values for new buildings and for old buildings are generally higher than those for middle-aged properties. This effect can be described, for example, by the increased renovation rate for old properties and the generally higher quality of new buildings. While the u-shape transformation seems reasonable in a parametric context, the effects shown in Figures 4.1 and 4.2 indicate that this transformation cannot be supported here. Hence, this transformation can only be seen as a rough approximation of the true underlying relationship. While we also see the effect that middle-aged properties tend to have a lower market value than the average valuation, the increase is much more significant for properties with newer construction years than for older buildings. In particular, for properties built between 2010 and 2020, we see that the increase is already almost exponential. This trend reflects the current high demand for new buildings in major German cities. Due to the lack of supply, people are currently willing to pay significantly higher prices for properties on the market. Another explanation lies in the sharp rise in land and construction costs. Both the acquisition of a plot of land and the construction of property have become significantly more expensive in Germany's metropolitan regions in recent years, which is also ultimately reflected in market values.

Figure 4.1: Condominiums – Year of construction

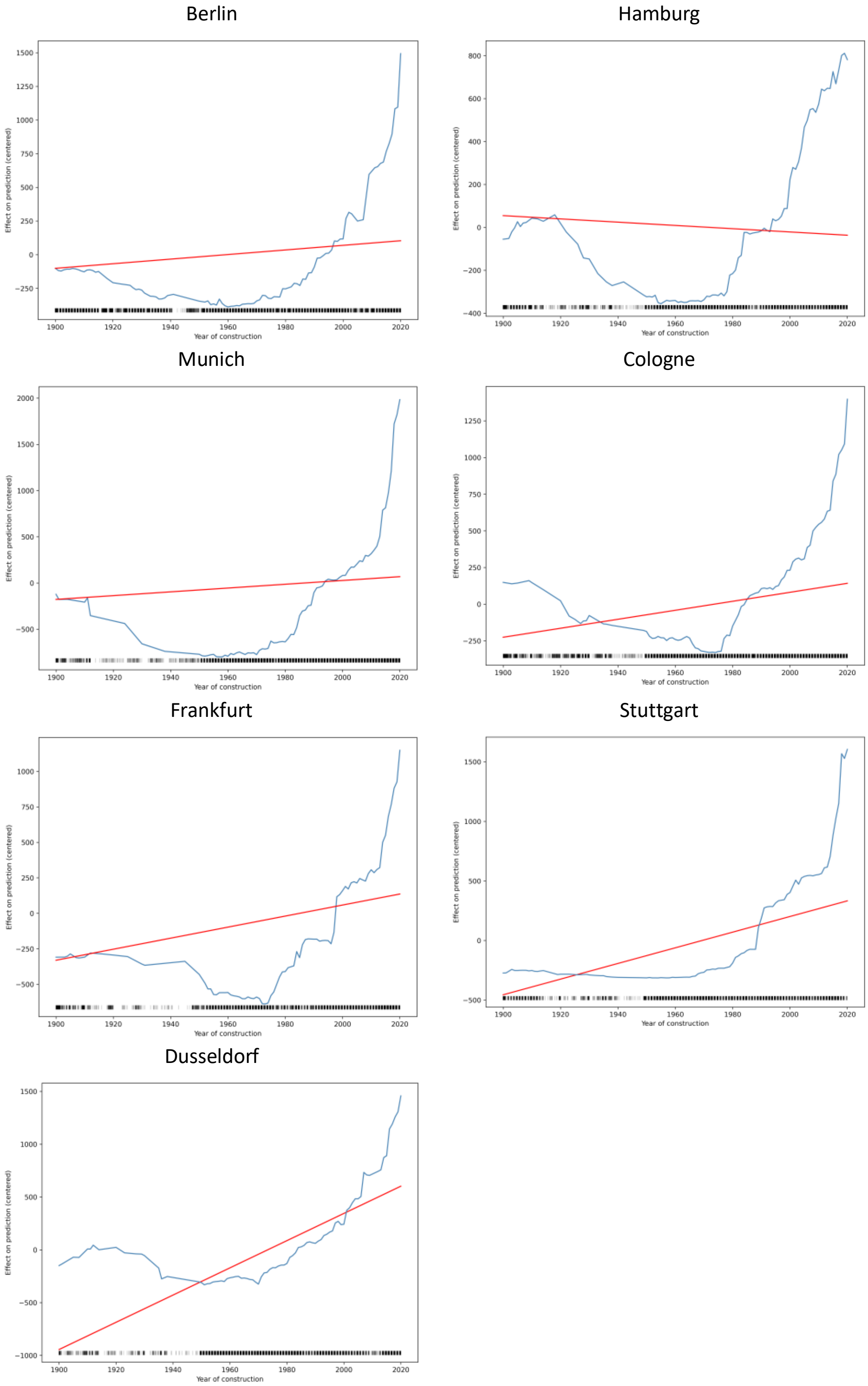


Figure 4.2: Single-family homes – Year of construction

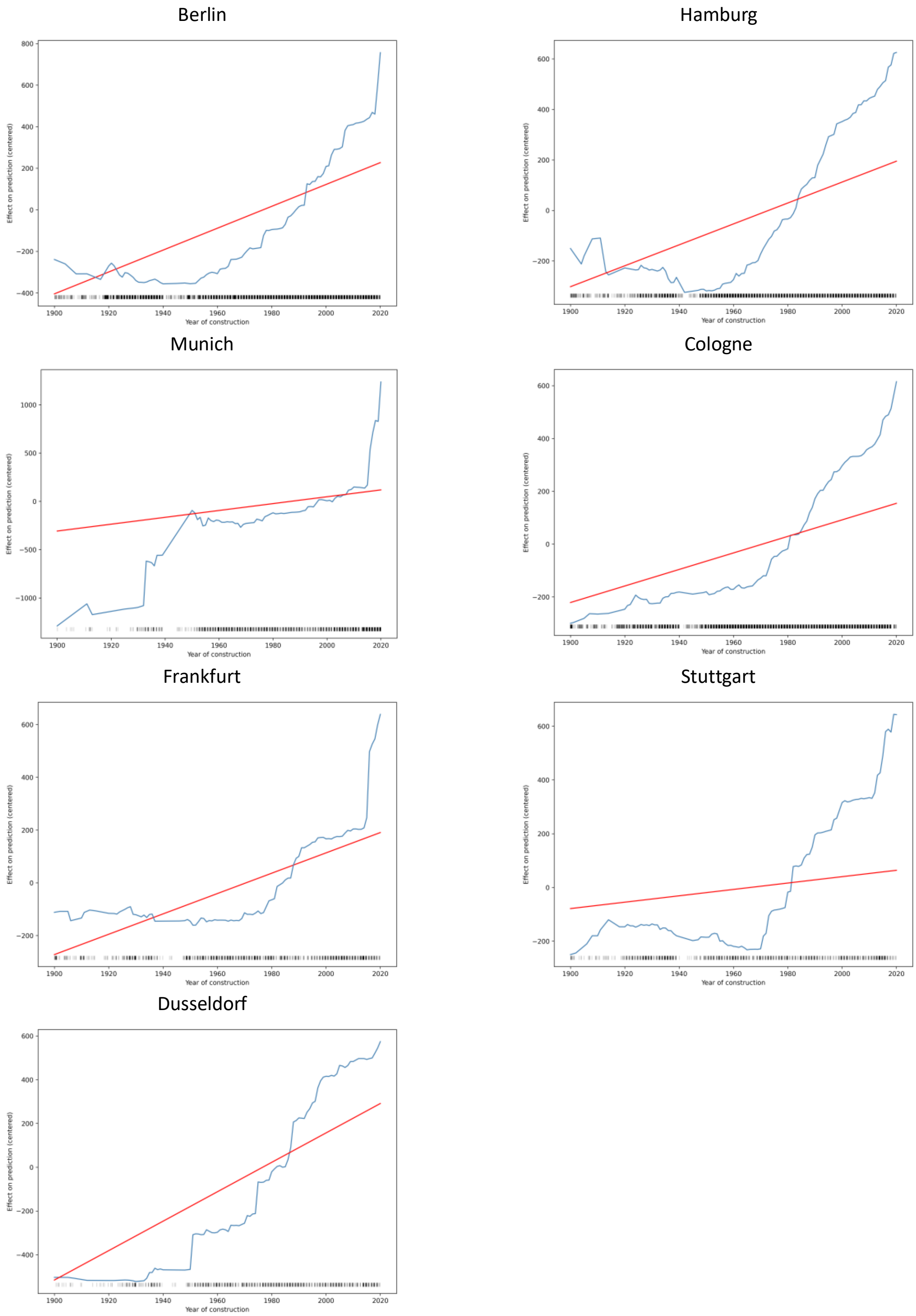
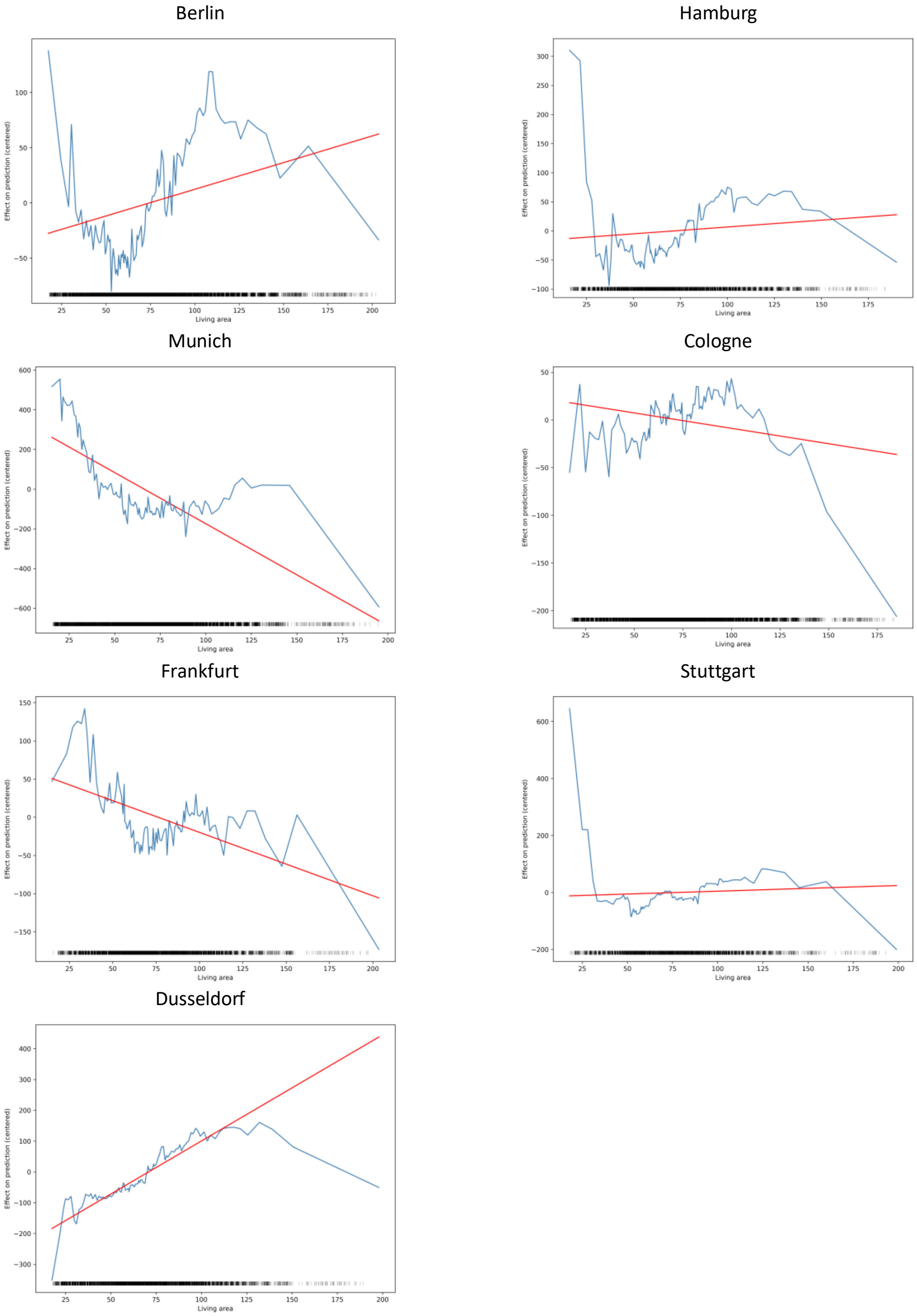


Figure 4.3: Condominiums – Living area



The effect of the living area on market values of condominiums and single-family homes can be seen in Figures 3 and 4. Focusing on the condominiums, for all Top-7 cities, a clear non-linear relationship is identified with the XGBoost, which would be very difficult to represent by parametric or semi-parametric models. As no recurring pattern is evident, the effects seem to differ in each city. The findings show that there is no generally applicable rule for the analyzed cities and that a well-known rule of thumb in the real estate industry "the larger the area, the lower the market value per square meter" does not hold for condominiums. The ALE plots clearly indicate that there are different patterns regarding the market values within the cities. For example, a high demand for small apartments in the cities of Berlin, Frankfurt, Hamburg, Munich, and Stuttgart is evident. In the cities of Dusseldorf and Cologne, on the other hand, this is not as pronounced. These results offer important implications for the real estate industry. The ALE plots support the analysis of which type of apartment sizes are in demand in which region and what prices can be achieved. Currently, such decisions are often still made on the basis of personal experience or purely descriptive market statistics. The combination of machine learning and ALE plots, on the other hand, enables an empirically valid and data-driven analysis. In contrast to condominiums, the effect of living area on the market value per square meter of single-family homes is homogeneous across all seven cities. Furthermore, the effect is almost linear and can be mapped by the basic OLS to a large extent. Overall, the effect shows a negative trend, which can most likely be attributed to the marginal cost effect. Major components of the costs of a single-family house (e.g., land area, development costs, etc.) are fixed to a certain extent and increase only noticeably as the size of the living area increases. In the case of larger houses, these costs are distributed over the additional square meters and lead to the negative trend shown in Figure 4.4. What also stands out is that the effect for single-family homes is larger on average, which indicates that the size of the property, compared to condominiums, is more important for houses.

Figure 4.4: Single-family homes – Living area

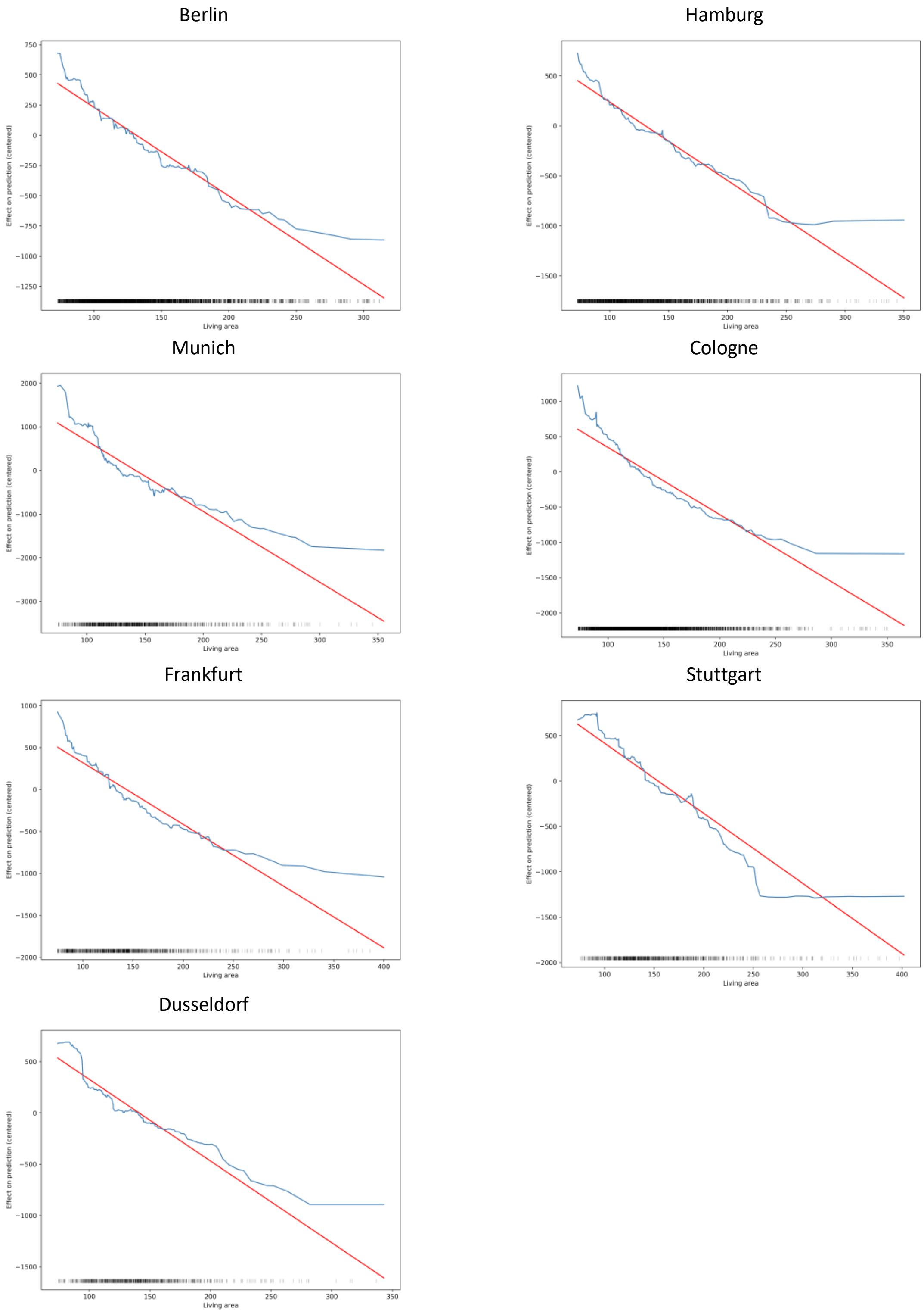


Figure 4.5: Single-family homes – Lot size

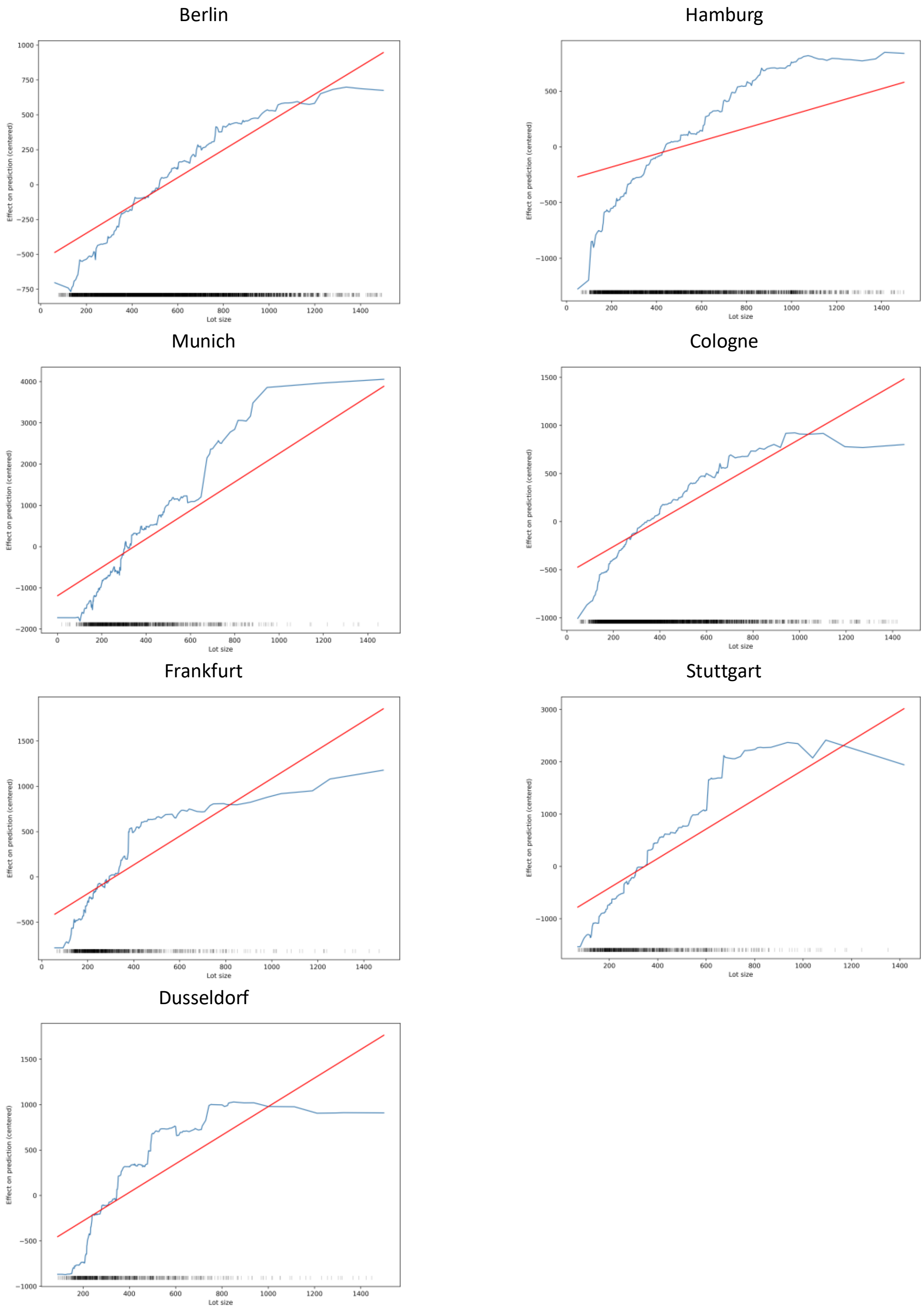


Figure 4.5 highlights the effect of lot size on the market value of single-family homes. In general, the effect is reasonably consistent across all Top-7 cities. The larger the lot size, the higher the market value per square meter of living area. This generally indicates that there is a higher demand for larger plots of land and that market values are rising due to a lack of supply, although a decreasing marginal utility can be seen for very large plots in all cities. While the basic OLS also assumes such a progression, the results of the XGBoost are more granular and thus more accurately reflect the actual effect. Differences between cities can be seen mainly in the strength of the effect. In Munich and Stuttgart, for example, the lot size seems to play a more important role than in Berlin or Hamburg. The results allow a more realistic picture to be drawn of the underlying market fundamentals, which ultimately allow real estate players to make better data-supported decisions.

As the results of the PFI show, the year of valuation was by far the most important feature. The effect of the valuation year is shown in Figure 4.6 for condominiums, and in Figure 4.7 for single-family houses. Since this variable is discrete, the next lower and next higher values are used as interval limits. The bars represent the size of the sample in each year, and the number is summarized with a second y-axis on the right of the plot. The red line again represents the results of a basic OLS. The results show that in all Top-7 cities, market values have risen sharply and constantly over the observation period. In principle, the OLS and XGBoost curves are relatively similar. However, it is apparent that the XGBoost identifies a stronger price increase for the last three years. In general, the demand for both condominiums and single-family homes has risen sharply in German metropolitan areas. Since supply is inelastic because of long development periods, this increase in demand leads to a dynamic rise in prices. Our findings show that this price increase also affects the market values of the properties. Property values have risen over the past few years, irrespective of their structural and locational features. This decoupling effect can be seen as quite critical, as the generally strongly rising prices can lead to speculation, which was also observable on the U.S. residential real estate markets before the Global Financial Crisis (Martin (2011)). In combination with a longer time series and other important macroeconomic features, the ALE plots could be used to conduct a more in-depth analysis and thus analyze key developments and drivers of real estate price bubbles. We consider this to be a promising and interesting area of research that should be pursued further.

Figure 4.6: Condominiums – Year of valuation

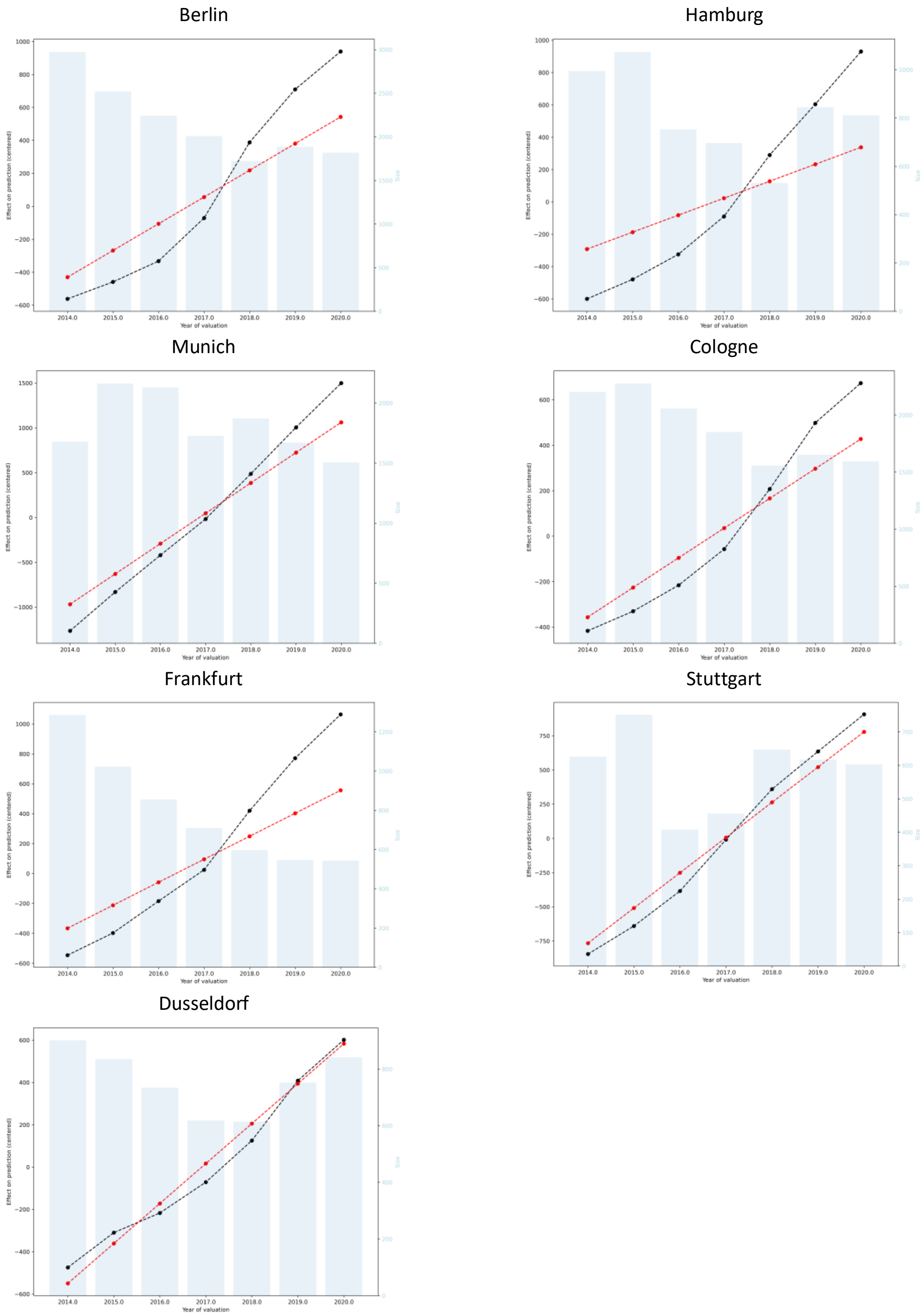
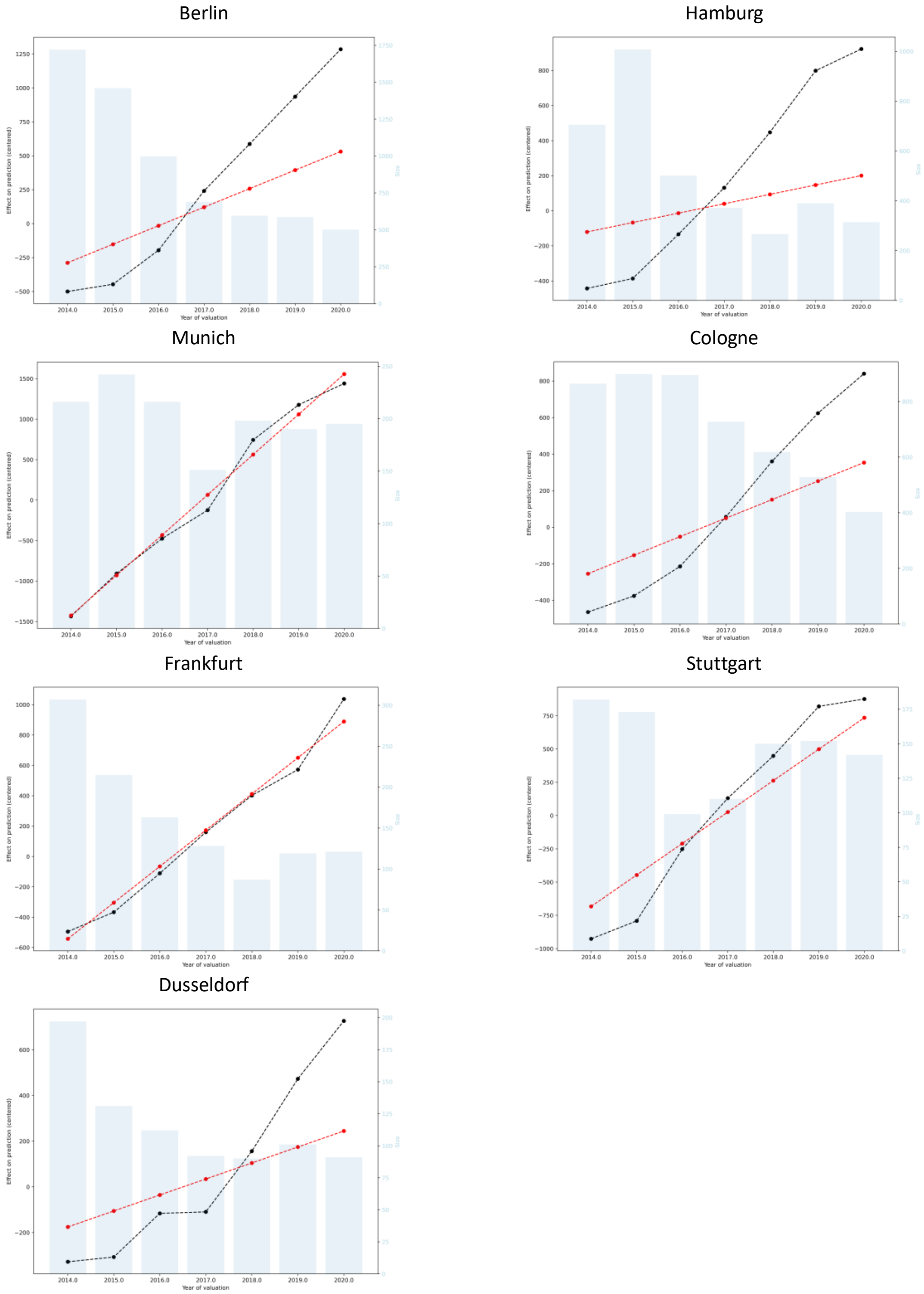


Figure 4.7: Single-family homes – Year of valuation



In addition to the temporal and structural characteristics, the permutation feature importance analysis shows that the location-related features of latitude and longitude play a significant role in predicting the market values per square meter for both condominiums and single-family homes. The effect of the latitude for condominiums is shown in Figure 4.8 and the effect of the longitude in Figure 4.9. In both figures, one can see that prices are rising sharply in certain regions – the city centers. An exception to this is the latitude of the city of Stuttgart. This is due to the unique location of Stuttgart, in a valley with a lot of industry in central locations and thus has its own geographical characteristics. Comparing the ALE plots of the XGBoost and the basic OLS, it can be clearly stated that a simple OLS cannot reproduce these non-linear locational effects. Looking at the impact of latitude and longitude on single-family houses in Figures 4.10 and 4.11, a clear difference can be seen. Not only is the effect much less pronounced, but the expensive regions are no longer in the center of the cities, which is not surprising, since the houses in these cities are located in the suburbs. In summary, the ALE plots of latitude and longitude can help to identify promising locations within the cities, and are therefore helpful for almost all players in the real estate industry. However, these results have to be interpreted with caution. There are several location-based features in our dataset. Besides latitude and longitude, there are the four micro-scores, which also describe the surrounding location of the properties. Furthermore, there are three socio-economic variables in the dataset, which are available at the zip code level and thus could also be seen as a proxy for location. To obtain the overall effect of the location on the price, these individual effects would have to be aggregated.

Figure 4.8: Condominiums – Latitude

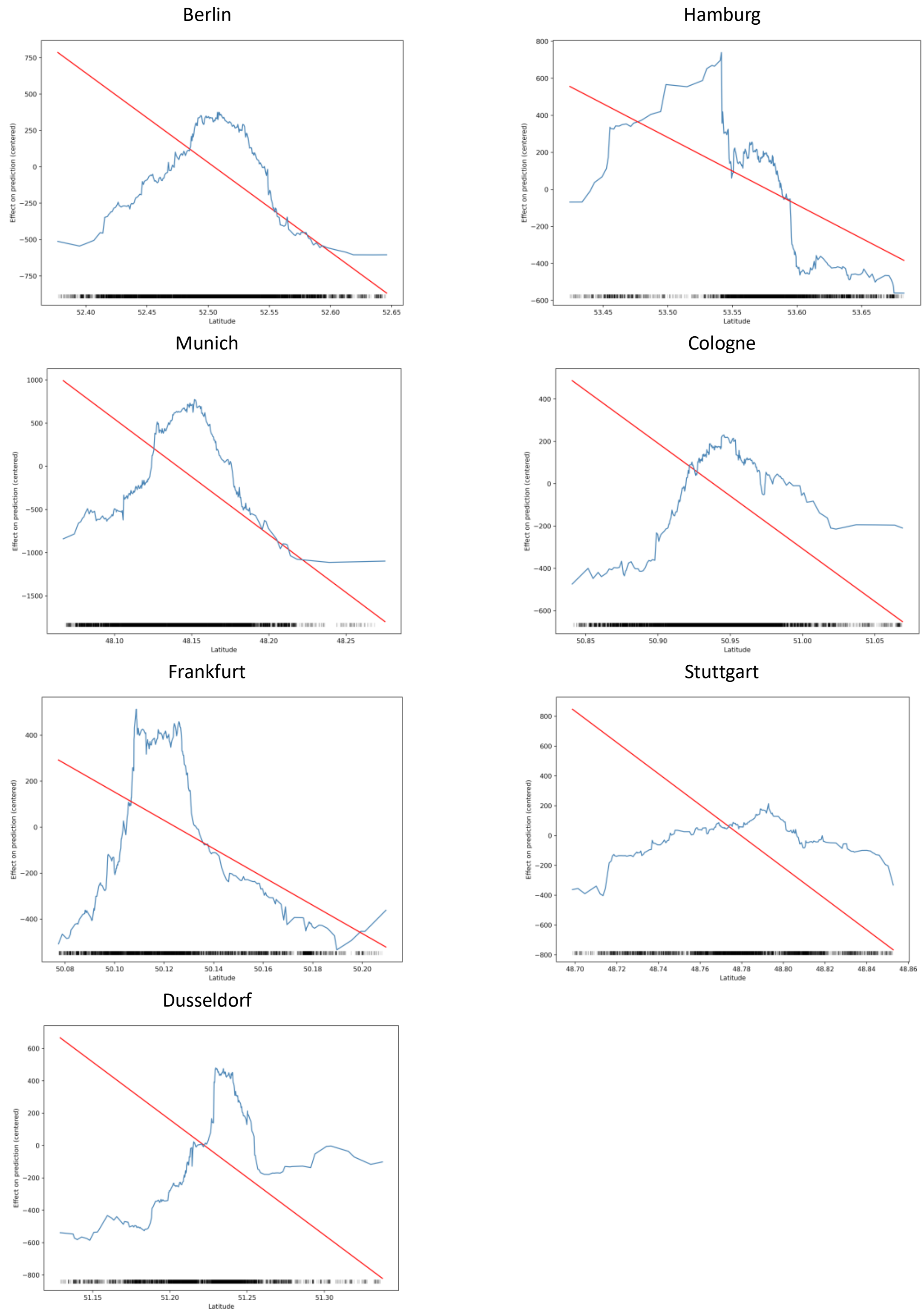


Figure 4.9: Condominiums – Longitude

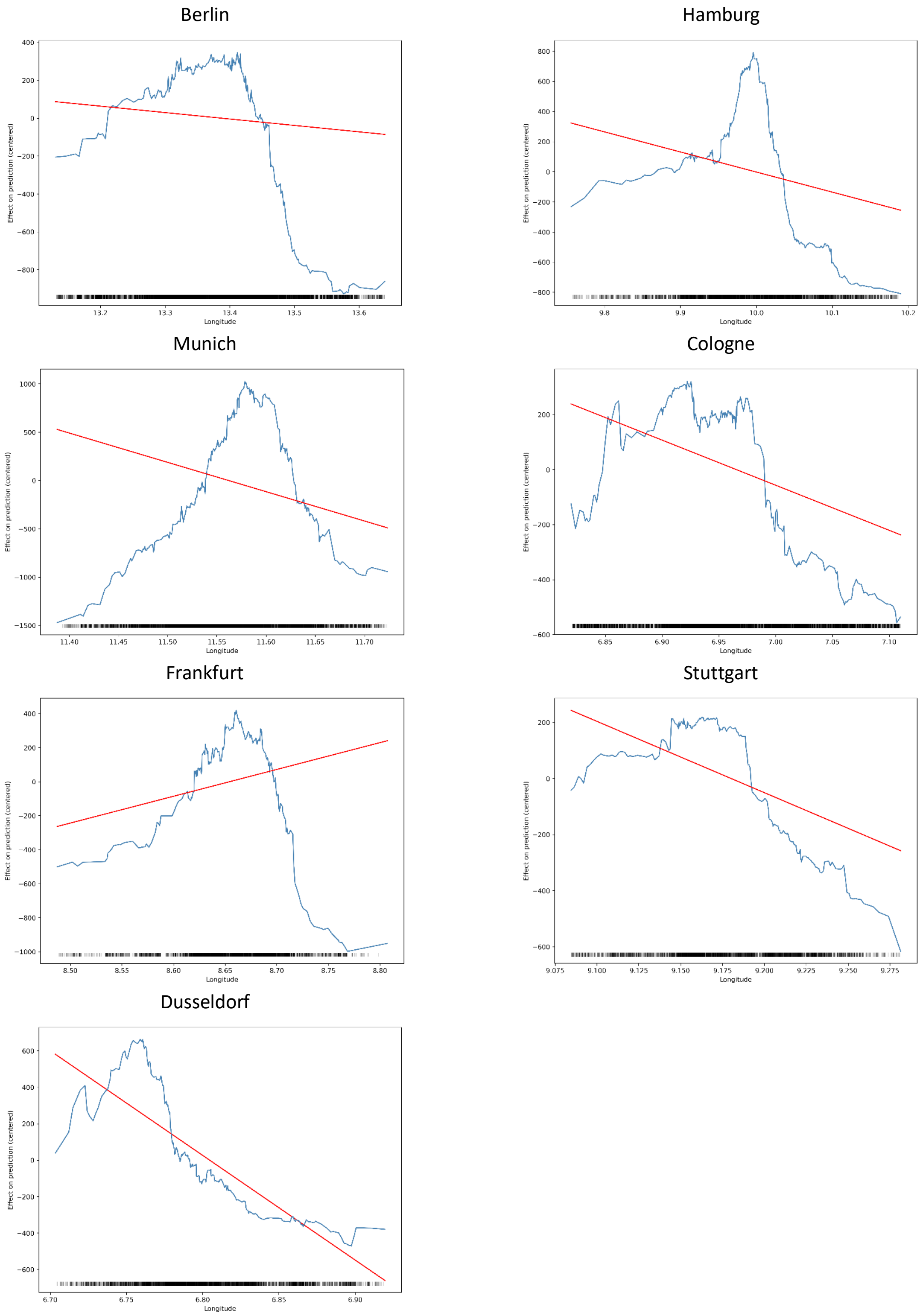


Figure 4.10: Single-family homes – Latitude

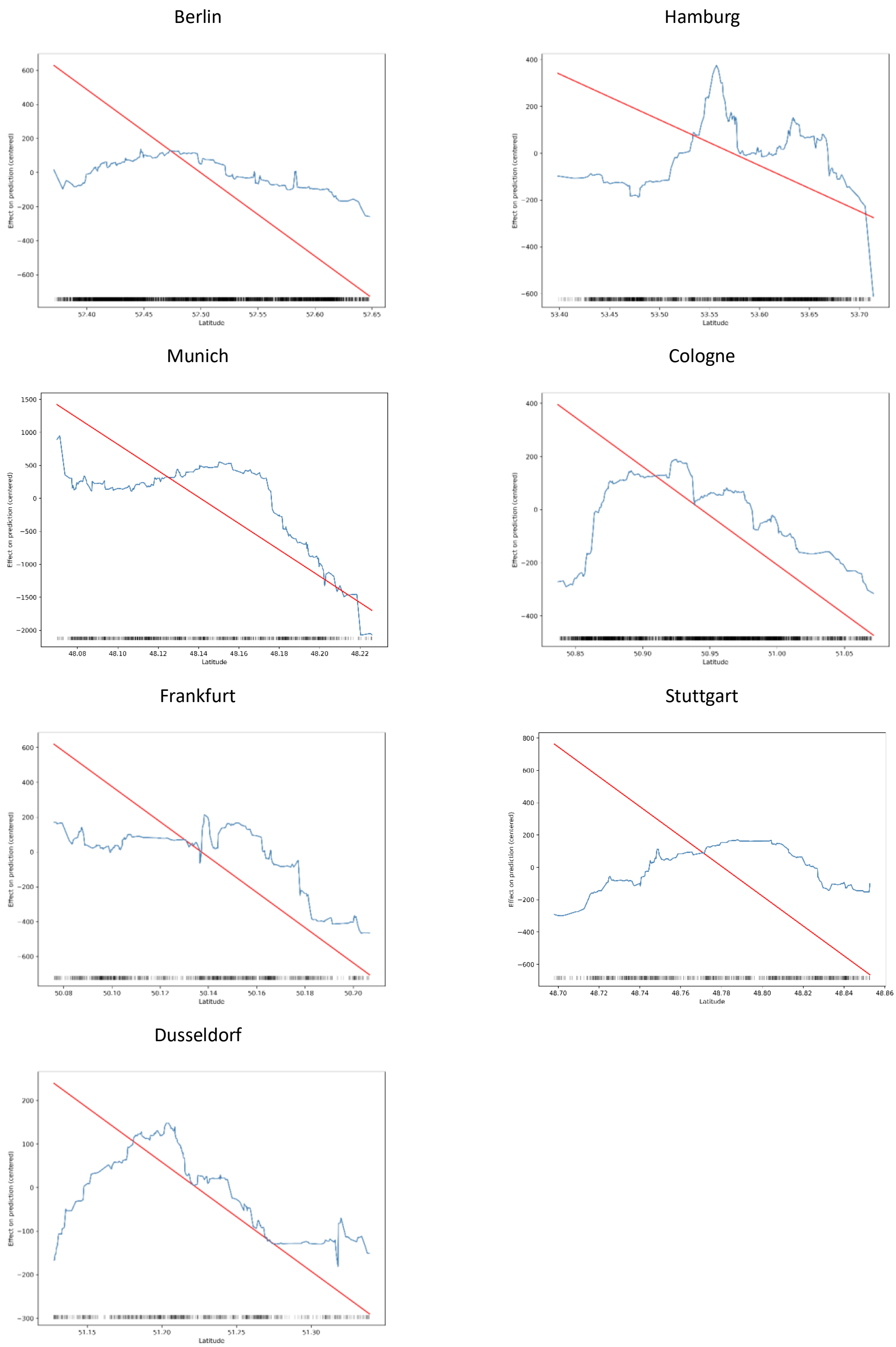
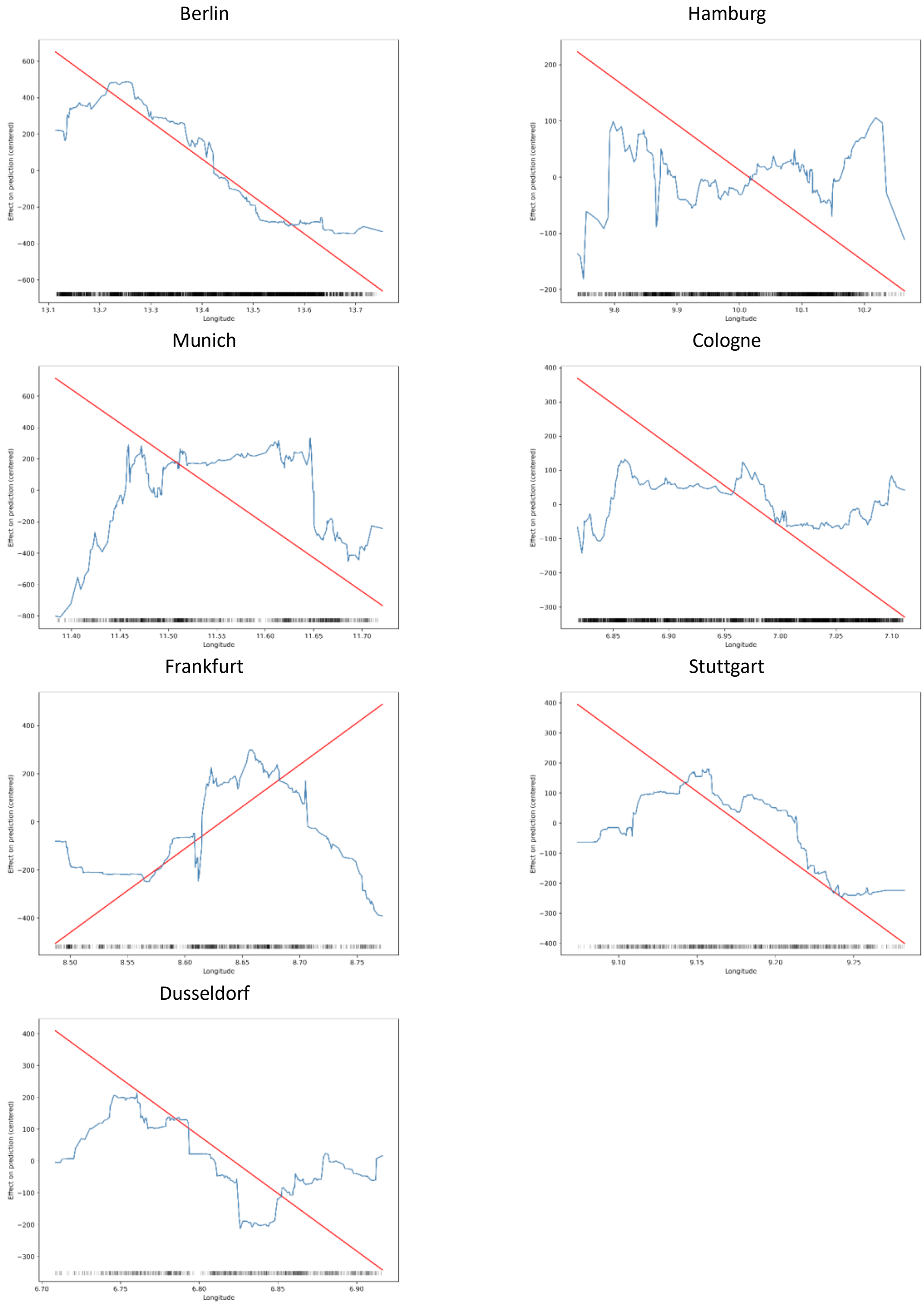


Figure 4.11: Single-family homes – Longitude



Finally, we consider the impact of the unemployment ratio on market values. Figure 4.12 shows the impact on condominiums and Figure 4.13 on single-family homes. Overall, the ALE plots of the XGBoost seem reasonable and are in line with the findings of other studies (see, e.g., Grum and Govekar (2016)). In all cities, the XGBoost identifies a downward trend in market values the higher the unemployment rate. Comparing the results of the XGBoost with the baseline OLS, one can see a large difference between the two graphs. In this context, the focus is on Stuttgart in particular, where positive effect of the OLS is assumed. Once again, the OLS is not able to capture the effects in a granular and comprehensive way. Decisions made on the basis of this flawed assumption can have far-reaching consequences and should be avoided. However, it bears repeating at this point that the results should also be interpreted with caution, as the unemployment rate can also serve as a simple proxy for the location of a property, due to its availability at the ZIP code level.

Figure 4.12: Condominiums – Unemployment ratio

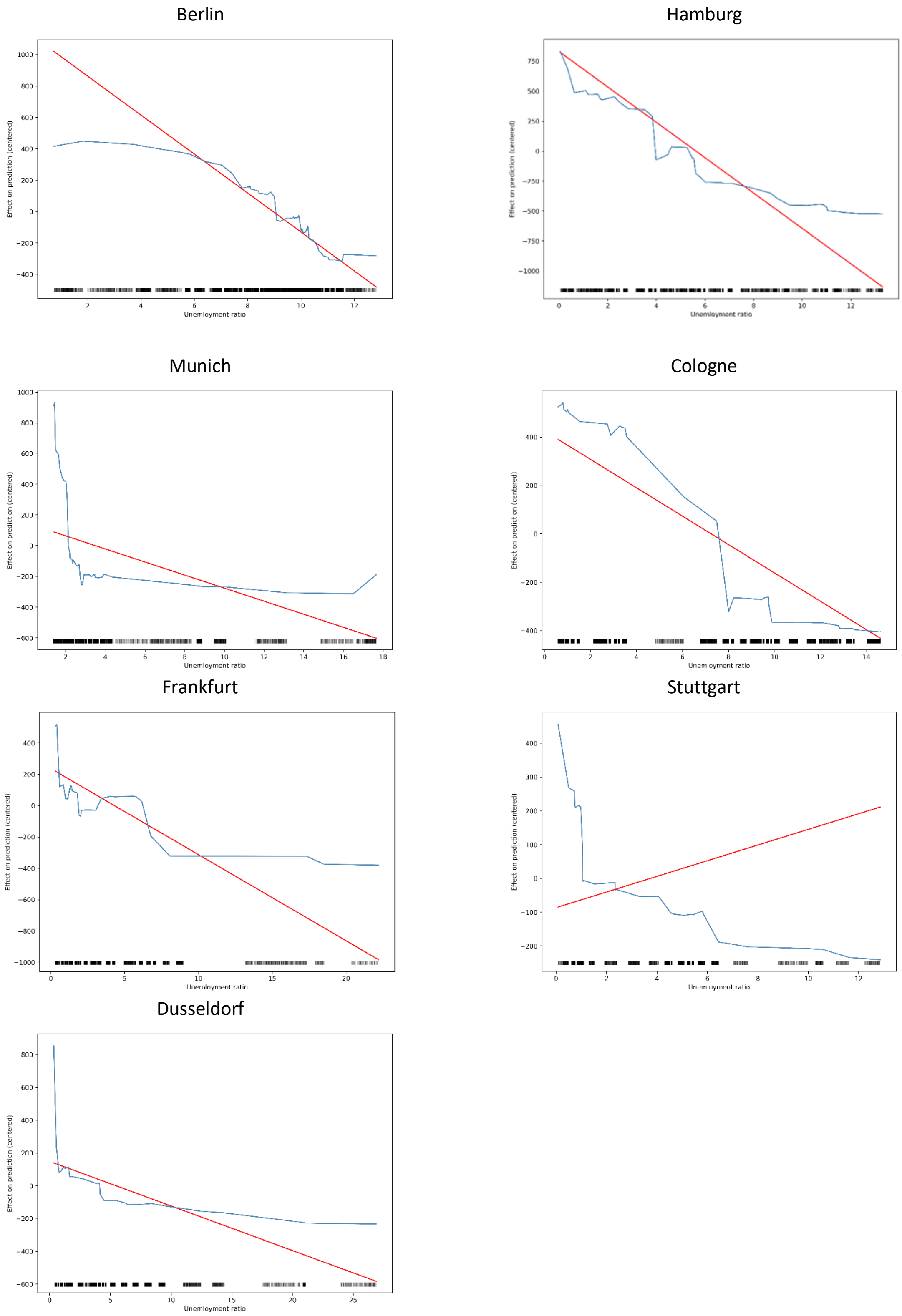
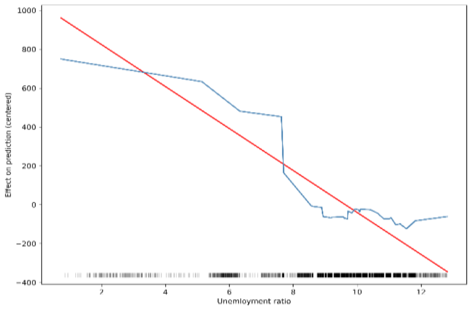
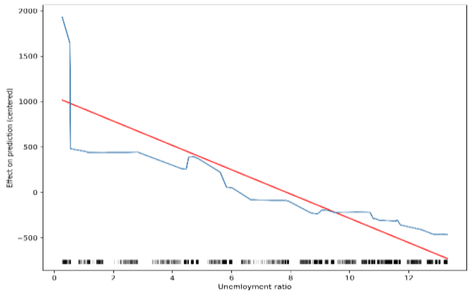


Figure 4.13: Single-family homes – Unemployment ratio

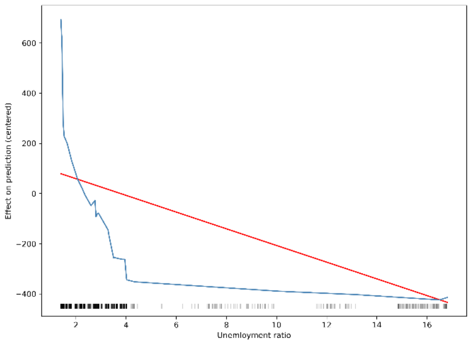
Berlin



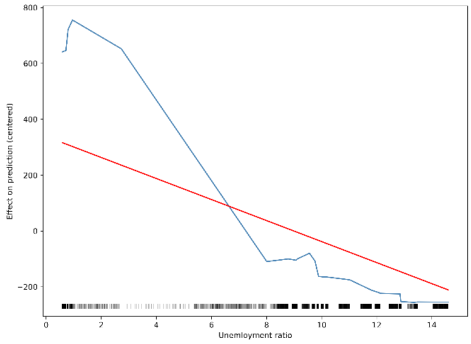
Hamburg



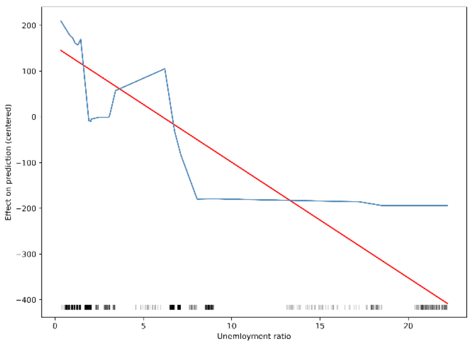
Munich



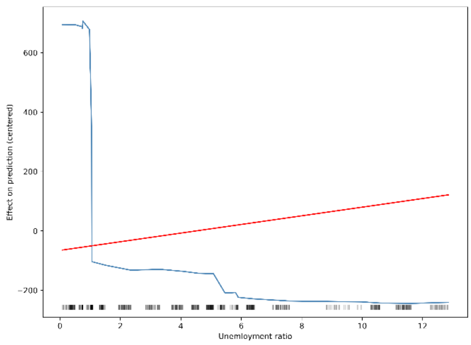
Cologne



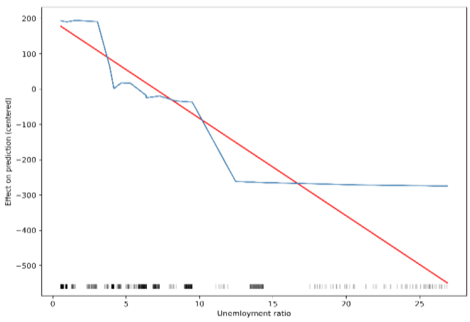
Frankfurt



Stuttgart



Dusseldorf



4.7 Conclusion

This study is intended to introduce XAI in a real estate context, and updates the existing literature with the application of ALE plots. Compared to the PDP plots, which are commonly used in real estate research to date, ALE plots can also handle dependent features and are therefore more suitable for real-estate-related problems. We use a dataset consisting of 61,763 condominiums and 19,403 single-family homes for the Top-7 cities of Germany to empirically apply our analysis. We are interested in identifying the most value-determining features of the two property subtypes, and then analyzing them separately with respect to existing non-linearities. We apply PFI to identify the most important features, and ALE plots to visualize their individual effects. As an underlying ML model, we use the XGBoost algorithm for a hedonic estimation of the properties' market values.

The PFI analysis shows that the same features play a predominant role for both condominiums and single-family homes. The valuation year is by far the most important feature. This finding is interesting, as the year of valuation feature is purely a factor describing the time and shows that the market values of the properties are mainly influenced by general market trends. In addition to the valuation year, a mix of structural, location-related and socio-economic features influence the market value of the properties. Among the structural features, the most important are the year of construction, the living area and the lot size. In terms of location features, the latitude and longitude are decisive in terms of market values. Socio-economically, we find that the unemployment ratio seems to play a partially important role.

To the best of our knowledge, we are the first to use ALE plots for visualizing individual effects on the market value, and we see that both non-linear and linear effects can be observed. In terms of year of construction, our results show that for both condominiums and single-family homes, the u-shaped transformation traditionally used for HPMs is not evident. Properties with newer construction years are valued much higher than is the case for old buildings. We can confirm that properties with middle age tend to have lower market values. The results for living area among condominiums are particularly interesting. The ALE plots show no clear trend here for the cities studied, but that this effect varies greatly and is clearly non-linear. For single-family homes, on the other hand, a linear trend can be observed for all cities. An approximately linear trend is also evident for the year of valuation feature. The analysis of the ALE plots of latitude and longitude shows that market values within the city can vary greatly, depending on the particular location. Our results show that market values for condominiums tend to be highest in the centers of cities, whereas values for single-family homes tend to be highest outside city centers. Both findings appear intuitive and are in line with the prevailing opinion within the real estate industry. The effect of the unemployment rate is also clearly non-linear and different across the cities analyzed. In general, however, the presumed negative influence is evident.

In summary, the ALE plots provide a deeper understanding of the fundamentals of real estate markets and either empirically confirm long-established rules of thumb or, as in the case of living area for condominiums, challenge them. Our results show that linear relationships indeed occur in the housing market. Here, parametric estimates can also provide valuable results. However, the analysis of the features year of construction, living area, lot size, latitude, longitude and unemployment ratio reveal non-linear effects. Therefore, non-parametric ML approaches seem to be the right choice. The ALE plots offer a way to represent these effects in a well-founded way and thus make an important contribution to the housing market literature. Both real estate research and practice can benefit from these results. The ALE plots provide a more in-depth analysis and thus examine key developments and drivers of real estate price bubbles, which offers an interesting area of research that should be pursued further. Model-agnostic methods are still a rather young field of research, but will play a major role in the acceptance of ML methods in the future, as they allow us to look into the "black box" of ML approaches and are thus an important tool in deciphering them. The tradeoff between explainability and model performance can thus be mitigated in the long run. However, further research is still needed before widespread use is possible.

4.8 Appendix

Appendix I – Micro Score

Our gravity model can be described using an activity function $f(A_p)$ and a distance function $f(D_{i,p})$:

$$A_{i,p} = \sum f(A_p) f(D_{i,p}).$$

$A_{i,p} \in [0,100]$ denotes the accessibility of point i for the POI p, whereby the activity function $f(A_p)$ specifies the relative importance of POI p, with $f(A_p) \in [0,1]$. $f(D_{i,p})$ measuring the travel time from point i to the POI p by using a non-symmetric sigmoidal distance function. The travel time was obtained for the selected POIs via Open Street Map, and normalized using the following function:

$$L(x) = \frac{K}{(1 + Qe^{0.5x})^{\frac{1}{v}}},$$

where $K, Q \in \mathbb{R}$ and $v \in \mathbb{R}^+$ are defined for all possible distances $x \in \mathbb{R}$. Furthermore, we have:

$$K = (1 + Q)^{1+v},$$

$$Q = v \cdot \exp(B \cdot x^*),$$

$$v = \frac{\exp(B \cdot x^*) - 1}{\ln(y_i) - 1},$$

where x^* denotes a feature specific point of inflection and y^* is 0.5.

Appendix II – Evaluation metrics at city level

Table 4.9: Evaluation metrics - City level

	XGBoost		OLS	
Metrics	Single-family homes	Condominiums	Single-family homes	Condominiums
		Berlin		
MAPE	0.1415	0.1431	0.1837	0.2311
MdAPE	0.0988	0.0984	0.1342	0.1741
PE(10)	0.5062	0.5059	0.3927	0.3077
PE(20)	0.7910	0.7752	0.6698	0.5593
R^2	0.7544	0.8052	0.6074	0.6192
		Hamburg		
MAPE	0.1505	0.1291	0.2039	0.1990
MdAPE	0.1047	0.0806	0.1466	0.1455
PE(10)	0.4805	0.5697	0.3595	0.3721
PE(20)	0.7710	0.8199	0.6450	0.6369
R^2	0.7245	0.7936	0.5288	0.6123
		München		
MAPE	0.1735	0.1051	0.2016	0.1718
MdAPE	0.0981	0.0618	0.1251	0.1233
PE(10)	0.5099	0.6559	0.4154	0.4163
PE(20)	0.7670	0.8772	0.7095	0.7104
R^2	0.6381	0.8079	0.5264	0.5734
		Köln		
MAPE	0.1232	0.1278	0.1587	0.2008
MdAPE	0.0878	0.0865	0.1180	0.1469
PE(10)	0.5530	0.5577	0.4381	0.3609
PE(20)	0.8220	0.8267	0.7217	0.6398
R^2	0.7256	0.8388	0.5672	0.6820
		Frankfurt		
MAPE	0.1571	0.1121	0.1866	0.2124
MdAPE	0.1041	0.0738	0.1222	0.1571
PE(10)	0.4816	0.6080	0.4351	0.3360
PE(20)	0.7640	0.8512	0.7070	0.6050
R^2	0.6639	0.8312	0.5431	0.6061

		Stuttgart		
MAPE	0.1513	0.1160	0.1859	0.1449
MdAPE	0.1033	0.0824	0.1337	0.1082
PE(10)	0.4851	0.5772	0.3720	0.4698
PE(20)	0.7778	0.8508	0.6845	0.7607
R^2	0.6869	0.8002	0.5518	0.6855
		Dusseldorf		
MAPE	0.1848	0.1334	0.2019	0.1907
MdAPE	0.1270	0.0930	0.1450	0.1438
PE(10)	0.4158	0.5296	0.3624	0.3622
PE(20)	0.6695	0.7921	0.6474	0.6440
R^2	0.4909	0.7793	0.4188	0.5867

4.9 References

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Bao, H. X., & Wan, A. T. (2004). On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data. *Real Estate Economics*, *32*(3), 487–507. https://doi.org/10.1111/j.1080-8620.2004.00100.x

Below, S., Beracha, E., & Skiba, H. (2015). Land Erosion and Coastal Home Values. *Journal of Real Estate Research*, *37*(4), 499–536. https://doi.org/10.1080/10835547.2015.12091427

Bogin, A. N., & Shui, J. (2020). Appraisal Accuracy and Automated Valuation Models in Rural Areas. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 40–52. https://doi.org/10.1007/s11146-019-09712-0

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial Dependence, Housing Submarkets, and House Price Prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), 143–160. https://doi.org/10.1007/s11146-007-9036-8

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions. *Journal of Real Estate Research*, *32*(2), 139–159. https://doi.org/10.2139/ssrn.1090147

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brunauer, W. A., Lang, S., Wechselberger, P., & Bienert, S. (2010). Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna. *The Journal of Real Estate Finance and Economics*, *41*(4), 390–411. https://doi.org/10.1007/s11146-009-9177-z

Cajias, M., & Freudenreich, P. (2018). Exploring the determinants of liquidity with big data – market heterogeneity in German markets. *Journal of Property Investment & Finance*, *36*(1), 3–18. https://doi.org/10.1108/JPIF-01-2017-0006

Can, A. (1990). The Measurement of Neighborhood Dynamics in Urban House Prices. *Economic Geography*, *66*(3), 254. https://doi.org/10.2307/143400

Carstens, R., Fang, Y., & Freybote, J. (2020). The Impact of Municipal Water Use Restrictions on the Pricing of Water-Sensitive Features in Single-Family Homes. *Journal of Real Estate Research*, *42*(4), 476–498. https://doi.org/10.1080/08965803.2020.1845553

Cheng, A. C. S., & Fung, M. K. (2015). Determinants of Hong Kong's Housing Prices. *Journal of Economics, Business and Management*, *3*(3), 352–355. https://doi.org/10.7763/joebm.2015.v3.208

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location. *Transportation Research Procedia*, *3*, 810–817. https://doi.org/10.1016/j.trpro.2014.10.067

Chun Lin, C., & Mohan, S. B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis*, *4*(3), 224–243. https://doi.org/10.1108/17538271111153013

Conway, D., Li, C. Q., Wolch, J., Kahle, C., & Jerrett, M. (2010). A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values. *The Journal of Real Estate Finance and Economics*, *41*(2), 150–169. https://doi.org/10.1007/s11146-008-9159-6

Des Rosiers, F., Lagana, A., & Theriault, M. (2001). Size and proximity effects of primary schools on surrounding house values. *Journal of Property Research*, *18*(2), 149–168. https://doi.org/10.1080/09599910110039905

Din, A., Hoesli, M., & Bender, A. (2001). Environmental Variables and Real Estate Prices. *Urban Studies*, *38*(11), 1989–2000. https://doi.org/10.1080/00420980120080899

Dubin, R. A. (1988). Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *The Review of Economics and Statistics*, *70*(3), 466. https://doi.org/10.2307/1926785

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2016). Price Variation in Waterfront Properties Over the Economic Cycle. *Journal of Real Estate Research*, *38*(1), 1–26. https://doi.org/10.1080/10835547.2016.12091435

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2018). Sinkholes and Residential Property Prices: Presence, Proximity, and Density. *Journal of Real Estate Research*, *40*(1), 41–68. https://doi.org/10.1080/10835547.2018.12091491

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies*, *43*(12), 2301–2315. https://doi.org/10.1080/00420980600990928

Fernández-Avilés, G., Mínguez, R., & Montero, J.-M. (2012). Geostatistical Air Pollution Indexes in Spatial Hedonic Models: The Case of Madrid, Spain. *Journal of Real Estate Research*, *34*(2), 243–274. https://doi.org/10.1080/10835547.2012.12091330

Fisher, A., Rudin, C., & Dominici, F. (2019). *All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously*. https://www.jmlr.org/papers/volume20/18-760/18-760.pdf

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Grum, B., & Govekar, D. K. (2016). Influence of Macroeconomic Factors on Prices of Real Estate in Various Cultural Environments: Case of Slovenia, Greece, France, Poland and Norway. *Procedia Economics and Finance*, *39*, 597–604. https://doi.org/10.1016/S2212-5671(16)30304-5

Gu, G., & Xu, B. (2017). Housing Market Hedonic Price Study Based on Boosting Regression Tree. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *21*(6), 1040–1047. https://doi.org/10.20965/jaciii.2017.p1040

Handy, S. L., & Clifton, K. J. (2001). Evaluating Neighborhood Accessibility: Possibilities and Practicalities. *Journal of Transportation and Statistics*, *4*(2). https://www.researchgate.net/profile/kelly-clifton/publication/235358397_evaluating_neighborhood_accessibility_possibilities_and_practicalities

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

Ho, W. K., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hoen, B., & Atkinson-Palombo, C. (2016). Wind Turbines, Amenities and Disamenitites: Astudy of Home Value Impacts in Densely Populated Massachusetts. *Journal of Real Estate Research*, *38*(4), 473–504. https://doi.org/10.1080/10835547.2016.12091454

Huang, Y., & Dall'erba, S. (2021). Does Proximity to School Still Matter Once Access to Your Preferred School Zone Has Already Been Secured? *The Journal of Real Estate Finance and Economics*, *62*(4), 548–577. https://doi.org/10.1007/s11146-020-09761-w

Hwang, J. W., Kuang, C., & Bin, O. (2019). Are all Homeowners Willing to Pay for Better Schools? – Evidence from a Finite Mixture Model Approach. *The Journal of Real Estate Finance and Economics*, *58*(4), 638–655. https://doi.org/10.1007/s11146-018-9658-z

Jauregui, A., Allen, M. T., & Weeks, H. S. (2019). A Spatial Analysis of the Impact of Float Distance on the Values of Canal-Front Houses. *Journal of Real Estate Research*, *41*(2), 285–318. https://doi.org/10.1080/10835547.2019.12091527

Kagie, M., & van Wezel, M. (2007). Hedonic price models and indices based on boosting applied to the Dutch housing market. *Intelligent Systems in Accounting, Finance & Management*, *15*(3-4), 85–106. https://doi.org/10.1002/isaf.287

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, *43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2021). Interpretable Machine Learning for Real Estate Market Analysis. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3835931

Lu, J., & Lee, Y. (2022). Spatial Effects of Air Pollution on the Housing Market: Evidence from South Korea. *Journal of Real Estate Research*, *44*(1), 131–150. https://doi.org/10.1080/08965803.2021.2008109

Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. *Housing Economics and Public Policy*, 67–89. https://doi.org/10.1002/9780470690680.ch5

Martin, R. (2011). The local geographies of the financial crisis: From the housing bubble to economic recession and beyond. *Journal of Economic Geography*, *11*(4), 587–618. https://doi.org/10.1093/jeg/lbq024

Mason, C., & Quigley, J. M. (1996). Non-parametric hedonic housing prices. *Housing Studies*, *11*(3), 373–385. https://doi.org/10.1080/02673039608720863

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, *12*(1), 134–150. https://doi.org/10.1108/JERER-08-2018-0035

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Metzner, S., & Kindt, A. (2018). Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties. *International Journal of Housing Markets and Analysis*, *11*(1), 73–100. https://doi.org/10.1108/IJHMA-02-2017-0018

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.

Nobis, C., & Kuhnimhof, T. (2018). *Mobilität in Deutschland − MiD: Ergebnisbericht*. https://elib.dlr.de/125879/

Pace, K. (1998). Appraisal Using Generalized Additive Models. *Journal of Real Estate Research*, *15*(1), 77–99. https://doi.org/10.1080/10835547.1998.12090916

Pace, R. K., & Hayunga, D. (2020). Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests. *The Journal of Real Estate Finance and Economics*, *60*(1-2), 170–180. https://doi.org/10.1007/s11146-019-09724-w

Peterson, S., & Flanagan, A. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. *Journal of Real Estate Research*, *31*(2), 147–164. https://doi.org/10.1080/10835547.2009.12091245

Powe, N. A., Garrod, G. D., & Willis, K. G. (1995). Valuation of urban amenities using an hedonic price model. *Journal of Property Research*, *12*(2), 137–147. https://doi.org/10.1080/09599919508724137

Sangani, D., Erickson, K., & Hasan, M. A. (2017). Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS).* IEEE. https://doi.org/10.1109/mass.2017.88

Schulz, R., Wersing, M., & Werwatz, A. (2014). Automated valuation modelling: a specification exercise. *Journal of Property Research*, *31*(2), 131–153. https://doi.org/10.1080/09599916.2013.846930

Simons, R. A., Seo, Y., & Rosenfeld, P. (2015). Modeling the Effects of Refinery Emissions on Residential Property Values. *Journal of Real Estate Research*, *37*(3), 321–342. https://doi.org/10.1080/10835547.2015.12091421

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), 1–44. https://doi.org/10.1080/10835547.2005.12090154

Stamou, M., Mimis, A., & Rovolis, A. (2017). House price determinants in Athens: a spatial econometric approach. *Journal of Property Research*, *34*(4), 269–284. https://doi.org/10.1080/09599916.2017.1400575

Stang, M., Krämer, B., Nagl, C., & Schäfers, W. (2021). *From Human Business to Machine Learning - Methods for Automating Real Estate Appraisals and their Practical Implications.* https://doi.org/10.2139/ssrn.3987051

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Turner, T. M., & Seo, Y. (2021). House Prices, Open Space, and Household Characteristics. *Journal of Real Estate Research*, *43*(2), 204–225. https://doi.org/10.1080/08965803.2021.1925498 van Wezel, M., Kagie, M., & Potharst, R. (2005). Boosting the accuracy of hedonic pricing models. *Report / Econometric Institute, Erasmus University Rotterdam*(EI 2005-50). https://repub.eur.nl/pub/7145/

Worzala, E., Lenk, M., & Silva, A. (1995). An Exploration of Neural Networks and Its Application to Real Estate Valuation. *Journal of Real Estate Research*, *10*(2), 185–201. https://doi.org/10.1080/10835547.1995.12090782

Wyman, D., & Mothorpe, C. (2018). The Pricing of Power Lines: A Geospatial Approach to Measuring Residential Property Values. *Journal of Real Estate Research*, *40*(1), 121–154. https://doi.org/10.1080/10835547.2018.12091490

Yang, J., Bao, Y., Zhang, Y., Li, X., & Ge, Q. (2018). Impact of Accessibility on Housing Prices in Dalian City of China Based on a Geographically Weighted Regression Model. *Chinese Geographical Science*, *28*(3), 505–515. https://doi.org/10.1007/s11769-018-0954-6

Zurada, J., Levitan, A., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, *33*(3), 349–388. https://doi.org/10.1080/10835547.2011.12091311

5 Conclusion

5.1 Executive Summary

The following provides a brief summary of the three papers from a concluding perspective.

Paper 1: Sentiment Analysis within a Deep Learning Probabilistic Framework – New Evidence from Residential Real Estate in the United States

The housing market is characterized by a heterogeneous stakeholder structure, low transparency, and a high information asymmetry, as described by Soo (2018). These facts make it a difficult and well-discussed task to understand the different drivers and influences on housing markets. Therefore, tailored research which is not limited to fundamental variables is particularly required. Nobel laureate Shiller (2000) notes that major market events in particular, such as price bubbles or turning points, depend besides the development of fundamental variables also on the expectations and behavior of the market participants.

Here connects the sentiment analysis, a subfield of behavioral finance. The goal is to identify current sentiment as a non-fundamental measure and thereby reveal important drivers of various markets, as described by Tetlock (2007). While much sentiment research has already been done in the financial literature, it has only gained increased traction in the real estate literature in the last decade. However, few studies have examined the housing market, for example, Hohenstatt et al. (2011), Dietzel et al. (2014), Walker (2014) and Soo (2018). In particular, the studies used the already established dictionary method.

This article develops a new method to measure sentiment in news media by combining the dictionary approach with ML. While each dictionary is designed for specific market participants, the ML method finds a way to combine them to create a tailored sentiment index. The main objective is to evaluate whether a sentiment index constructed from media text offers additional explanatory power for predicting monthly changes in the S&P Case-Shiller Home Price Index.

Using a unique data set of approximately 25,000 headlines and abstracts of real-estate related Wall Street Journal articles, this study contributes to the literature in three ways. First, the study introduces an innovative methodology for the consistent labelling of text sequences and delivers some initial evidence that combining different dictionaries has a larger and statistically significant impact on the housing market compared to single wordlists. Second, a novel use of ML in the housing market is provided and its capability to account for heterogeneous expectations in this unique market is revealed. Third, this study is not only limited to titles, but extends research in the housing market by also using abstracts, thereby demonstrating that the new methodology is superior for both types of news media.

Paper 2: From Human Business to Machine Learning – Methods for Automating Real Estate Appraisals and their Practical Implications

Due to the progressing digitization, the application of machine learning methods is also in focus in the field of real estate valuation. To support appraisers in their daily business, and not only for reasons of potential cost savings, so called AVMs have been established in recent years. Their achievement is also supported by the increasing data availability and quality, but also by the rising GPU power. AVMs already achieve good results by using for example simple hedonic models for calculating the market value. However, the choice of underlying algorithms is under discussion. Newly designed methods of ML are not yet established on the market and regulators in particular have expressed concerns about their use. Therefore, this study provides an automation of the traditional valuation approach – the comparison sales method - by applying various filters and similarity functions and compares these results to statistical and machine methods on a national-wide level for Germany.

Provided by a large German banking group, the analysis is based on a data set of 1,212,546 residential properties across Germany. The data originates from valuations and transactions of standard residential real estate lending and is collected between 2014 and 2020. A total of four different algorithms are used to value standard residential properties. Firstly, the automated comparison sales method replicates the traditional valuation methods. The hedonic pricing models include the OLS and the GAM. The ML methods are represented by the XG-Boost.

This study contributes to the regulatory debate of using AVMs in a theoretical and practical context by comparing traditional methods with a machine learning method. The results suggest that the use of ML approaches in particular is beneficial for AVMs. In a national comparison, this method most often shows the best results. However, based on the area-wide data set it becomes clear that, especially in rural areas, linear models also achieve good results and partly perform better than the ML approach. These results indicate that AVM construction does not depend on one single algorithm but asks for a tailored mix of different models to achieve the highest valuation accuracy, depending on the data availability and quality.

Paper 3: Explainable AI in a Real Estate Context – Exploring the Determinants of Residential Real Estate Values

Understanding market drivers is one of the most important challenges for understanding market mechanisms. This also holds for real estate values, which are influenced by a variety of different influences and drivers. While linear models are already well established in literature, in recent years several studies find non-linearity within the features of real estate values, for example Chun Lin and Mohan (2011), Kok et al. (2017) and Mayer et al. (2019) and Lorenz et al. (2021). Therefore, and as the results of Paper 2 have also shown, ML algorithms for understanding market structures and effects on property values are promising. However, ML algorithms are still considered to be black boxes, since it is not explainable how the algorithms get from the input to the output. The application of XAI might ensure an increasing acceptance of machine learning methods and represents a rather young field of research in real estate. The analysis is based on a data set of 81,166 residential properties in the Top-7-cities of Germany. The data originates from valuations and transactions of standard residential real estate lending by a large German banking group and was collected between 2014 and 2020. This study compares two XAI methods, namely Permutation Feature Importance (PFI), introduced by Breiman (2001), and Accumulated Local Effects (ALE) plots, derived by Apley and Zhu (2020), across Germany.

Using PFI and ALE plots, the computing process behind eXtreme Gradient-Boosting (XG-Boost) is revealed to understand which features influence the market value of properties and in which direction these effects work, i.e., whether the market values are influenced positively or negatively. Furthermore, potential non-linearity of the effects can be exposed. The study contributes to the real estate literature by increasing the transparency of ML methods and provides important insights in the drivers of the housing markets, which is relevant for various different stakeholders, such as appraisers, investors, and homeowners.

5.2 Final Remarks and Outlook

Due to the special characteristics of housing markets, their analysis and understanding of their mechanisms is a great challenge. On the one hand, housing is a basic human need on which a large part of a society's social stability depends. On the other hand, the residential asset class ties up a high investment volume for private and professional investors and plays a major role from a regulatory and financial stability perspective. Regardless of the motivation that drives researchers, analysts, politicians and other stakeholders, understanding the mechanisms and drivers in the housing market is highly important for many reasons. The incorporation of machine learning, which already has a history of nearly 80 years, can provide support.

This dissertation involves three main areas in which by increasing the transparency of housing markets using ML is studied in more detail. First, sentiment analysis is conducted to measure the expectations of various market participants. Concerning the different stakeholders of the housing market, it can be assumed that their expectations and sentiment differ from each other. A major challenge in sentiment analysis is the conversion of human language with all its facets into a computer-understandable sequence of data. The present work shows that the use of machine learning algorithms can support the processing of the complex correlations and effects within the text data and thus make them accessible for fine-grained market analysis. The newly designed sentiment index using a combination of different wordlists might also be supportive for further research, for example in detecting relevant turning points in the real estate cycle or for identifying price bubbles. Furthermore, the approach is not limited to the use of real estate and can also provide interesting findings for other research areas, where the expectations and mood of different stakeholders are relevant.

Second, in addition to the use of ML in the area of behavior finance, the influence of fundamental data on the housing market also stands in focus of this dissertation. Investigating the effects of different features on the market value, a comprehensive comparison of different approaches to automate property valuation for the German housing market is conducted. As the results show, the application of ML provides promising results and achieves a high accuracy in the valuation of standard residential properties. Again, the reason for this lies primarily in the efficient processing of complex relationships in the data. Especially in regions with high housing data availability, efficient algorithms are a beneficial support for the valuation of real estate. From a regulatory perspective, the application of AVMs based on ML algorithms should therefore be further discussed in order to find a reasonable combination of human knowledge and machine learning in practice. Best possible estimation results and the integration of all relevant data and information can avoid mispricing, reduce measurement errors and thus improve stability in the financial and housing market.

Third, to further support this process, the still rather young research field of XAI opens up new possibilities. The use of these methods makes it possible to obtain an approach to explaining the processes behind ML algorithms. Thus, these tools are not only of high importance for regulatory purposes. They allow the analysis of fundamentals relevant to the housing market and show the direction of the effects in which they operate. This knowledge is of interest to many different stakeholders who are concerned with the drivers of the real estate market. Among other things, fine-grained market analyses can be carried out from an investment perspective, looking at the question which characteristics of real estate are particularly in demand in local markets. Furthermore, price mechanisms can be analyzed, and thus, for example, the effect of political measures on the market can be monitored.

Overall, this dissertation shows that the use of ML algorithms in different areas of housing market analysis provides forward-looking results that pave the way for further research in this area. Improved data quality and increasing computing power are driven by the trends of digitization, Big Data and ML, forcing researchers and practitioners to address their use in the best possible way.

5.3 References

Apley, D.W. and Zhu, J. (2020), “Visualizing the effects of predictor variables in black box supervised learning models”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 82 No. 4, pp. 1059–1086.

Breiman, L. (2001), “Random Forests”, *Machine Learning*, Vol. 45 No. 1, pp. 5–32.

Chun Lin, C. and Mohan, S.B. (2011), “Effectiveness comparison of the residential property mass appraisal methodologies in the USA”, *International Journal of Housing Markets and Analysis*, Vol. 4 No. 3, pp. 224–243.

Dietzel, M.A., Braun, N. and Schäfers, W. (2014), “Sentiment-based commercial real estate forecasting with Google search volume data”, *Journal of Property Investment & Finance*, Vol. 32 No. 6, pp. 540–569.

Hohenstatt, R., Kaesbauer, M. and Schäfers, W. (2011), “"Geco" and its Potential for Real Estate Research: Evidence from the U.S. Housing Market”, *Journal of Real Estate Research*, Vol. 33 No. 4, pp. 471–506.

Kok, N., Koponen, E.-L. and Martínez-Barbosa, C.A. (2017), “Big Data in Real Estate? From Manual Appraisal to Automated Valuation”, *The Journal of Portfolio Management*, Vol. 43 No. 6, pp. 202–211.

Lorenz, F., Willwersch, J., Cajias, M. and Fuerst, F. (2021), “Interpretable Machine Learning for Real Estate Market Analysis”, *SSRN Electronic Journal*.

Mayer, M., Bourassa, S.C., Hoesli, M. and Scognamiglio, D. (2019), “Estimation and updating methods for hedonic valuation”, *Journal of European Real Estate Research*, Vol. 12 No. 1, pp. 134–150.

Shiller, R.J. (2000), “Measuring Bubble Expectations and Investor Confidence”, *Journal of Psychology and Financial Markets*, Vol. 1 No. 1, pp. 49–60.

Soo, C.K. (2018), “Quantifying Sentiment with News Media across Local Housing Markets”, *The Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Tetlock, P.C. (2007), “Giving Content to Investor Sentiment: The Role of Media in the Stock Market”, *The Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.

Walker, C.B. (2014), “Housing Booms and Media Coverage”, *Applied Economics*, Vol. 46 No. 32, pp. 3954–3967.