KWARTALNIK NAUK O PRZEDSIĘBIORSTWIE

1/2024

Chong Dae Kim Ph.D.
University of Applied Science
Cologne, Germany
Faculty of Process Engineering,
Energy and Mechanical Systems
e-mail: chong.kim@th-koeln.de

Nils Bedorf
University of Applied Science
Cologne, Germany
Faculty of Information,
Media and Electrical Engineering
e-mail: nils.bedorf@th-koeln.de

Predicting housing sale prices in Germany by application of machine learning models and methods of data exploration

Przewidywanie cen mieszkań w Niemczech z wykorzystaniem modeli uczenia maszynowego i metod eksploracji danych

Keywords:
machine learning, big data, sale price prediction, Germany, economics, real estate dataset

Abstract: The prediction of real estate prices is a popular problem in the field of machine learning and often demonstrated in literature. In contrast to other approaches, which regularly focus on the US market, this paper investigates the biggest, German real estate dataset, with more than 1.5 million unique samples and more than 20 features. In this paper we implement and compare different machine learning models in respect to performance and interpretability to give insight in the most important properties, which contribute to the sale price. Our experiments suggest that the prediction of sale prices in a real-world scenario is achievable yet limited by the quality of data rather than quantity. The results show promising prediction scores but are also heavily dependent on the location, which leaves room for further evaluation.

Słowa kluczowe:
uczenie maszynowe, big data, prognozowanie cen, Niemcy, ekonomia, zbiór danych o rynku nieruchomości

Streszczenie: Przewidywanie cen nieruchomości jest popularnym problemem w dziedzinie uczenia maszynowego i często przedstawianym w literaturze. W przeciwieństwie do innych podejść, które koncentrują się na rynku amerykańskim, niniejszy artykuł bada największy niemiecki zbiór danych dotyczących nieruchomości, zawierający ponad 1,5 mln unikatowych próbek i ponad 20 cech. W tym artykule wdrażamy i porównujemy różne modele uczenia maszynowego pod względem wydajności i możliwości interpretacji, aby uzyskać wgląd w najważniejsze

atrybuty mieszkania, które przyczyniają się do jego ceny sprzedaży. Nasze eksperymenty sugerują, że przewidywanie cen jest osiągalne, ale ograniczone przez jakość danych, a nie ich liczbę. Wyniki pokazują obiecujące wyniki predykcji, ale są również silnie zależne od lokalizacji, co pozostawia miejsce na dalsze badania.

JEL:
C4, C5, C52, C53, C55

Introduction

In the research field of machine learning, data scientists and big companies strive to solve the problems of our time by applying new algorithms to improve existing solutions even further. Significant advancements in terms of computing capabilities in the last decade unfolded new possibilities to approach data related challenges in various science related topics. There has never been a time before, where a broad audience, students and professionals alike, had freely access to advanced frameworks to design and use algorithms as well as the necessary hardware.

One of these problems lies within the space of economics, where new methods to efficiently predict real estate prices could be of immense use for the common user as well as self-employed agents and companies [Pow, Janulewicz, Liu, 2014]. While there already exist methods like Linear Regression to make such predictions the question remains if new machine learning algorithms may lead to performance improvements and further insights into the economic dynamics and relationships between distinct features in the real estate market.

In this paper we focus on an extensive approach to analyze and enhance a german real estate dataset with the goal to predict sale prices for the german housing market. This topic can be considered important since related work for the german market has not been published yet. Instead, current implementations and experiments usually focus on the american market [Wu, 2017; Park, Bae, 2015] on a much smaller scale and other countries [Viktorovich, Aleksandrovich, Leopoldovich, Vasilevna, 2018]. To predict our target variable, we use a variety of state-of-the-art algorithms, which include gradient boosting techniques [Gonzalez, Farcia et al., 2020; Géron, 2019], ensemble learning [Zhang, Ma, 2012], different regression algorithms, KNN and SVMs [Géron, 2019].

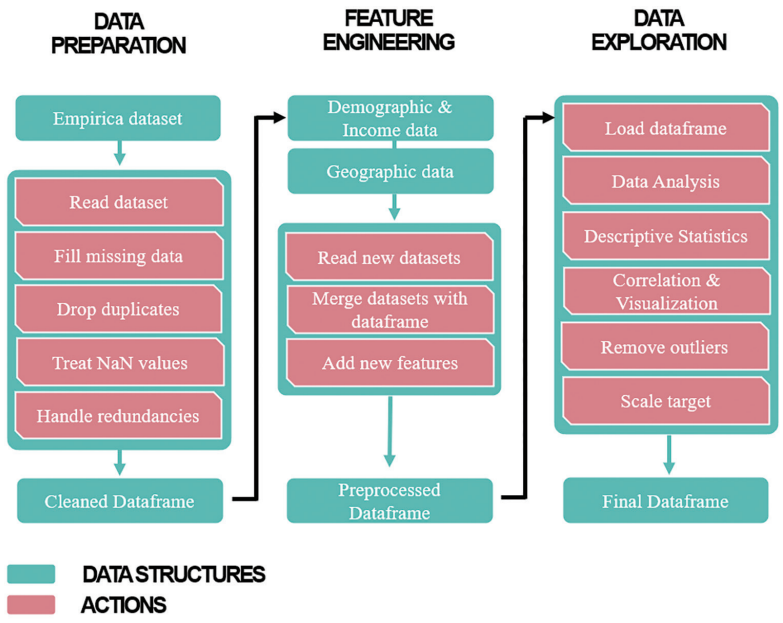
The first challenge is to comprehend the dataset itself, while finding additional data sources to complement it. This was ultimately achieved with help of the Federal Statistical Office in Germany and various other sources, which will be mentioned in chapter 3. It should also be mentioned that getting information in general about the real estate market (or anything) in Germany is far more difficult than in other countries, because of strict data protection laws [Bundesamt für Justiz, 2019], which make this dataset unique.

The next step is to clean the data without losing significant amounts of information. Preserving the integrity of the dataset is an important step, since the prediction should reflect the vast majority of all samples. A grid search over several parameter configurations and algorithms is performed and predictions for multiple regions are made, in order to distinct between locations with bad and good predictions. This is also done because the data is not distributed homogenously in Germany. While the results prove that predictions are possible and new approaches outperform classic regression, they also show that our data lacks information, which may contribute to better results in the future.

Overview

To give insight in the general procedure Fig. 1 shows a brief overview. On the step of "Data preparation" the dataset, which is provided by "Empirica ag" is read in.

Figure 1. Flowchart of data processing



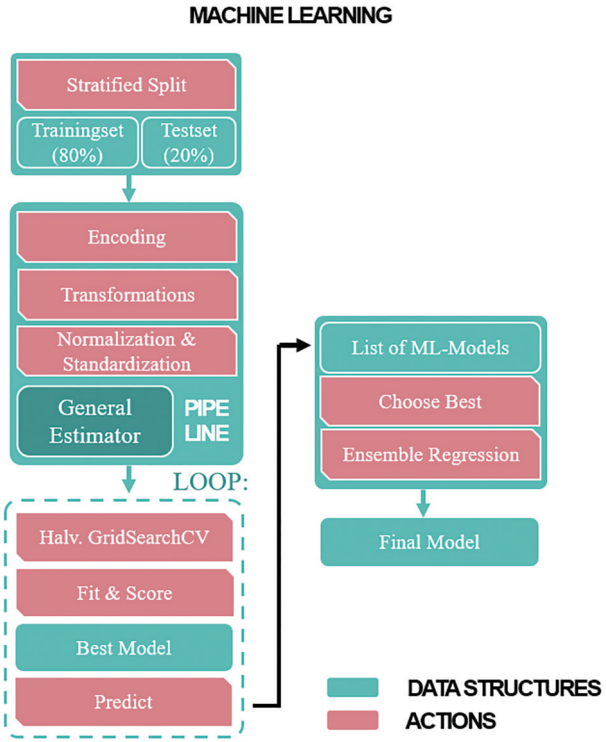
Source: Bedorf [2021].

Missing features are, if possible, filled in with data from other samples, duplicates are removed, and NaN values approximated or removed as well. Redundancies within categorical features are simplified to reduce dimensionality in future steps.

After the dataset is cleaned, we can add new features, based on other data. We use postcodes and counties as key to connect additional features to the data frame, for example: Geographic location (x, y) in meters, Income per household, Percentage of pensioners, population density, age of real estate, etc.

The pre-processed data frame will then be analyzed and visualized in terms of value distribution. To get some insight into feature dependencies the correlation matrix is computed, and unnecessary features are dropped. Outlier removal is a critical step, which is performed visually and by means of statistics in correspondence to the normalized target vector. This needs to be done before computation, because the target is not affected by transformations which are applied in the transformation pipeline.

Figure 2. **Machine Learning Pipeline**



Source: Bedorf [2021].

The final data frame as pictured in Fig. 2 is then split in such a way, that each geographic region is equally represented and fed into a pipeline. All categorical features are then binary or one-hot encoded. We also fit a Target Encoder on the training set before it is applied to the whole data set. The numerical features are transformed using

the "yeo-yohnson" [Yeo, Johnson, 2000, pp. 954–959] method, numpy log1p function or left unedited, depending on their skewness.

The data is then normalized and standardized. A general estimator class is implemented, which can take any algorithm as argument to perform a halving CV gridsearch over a arbitrary set of algorithms and parameter configurations. This makes the training process almost autonomous. The best cross-validated models are used for the prediction on the test set. The models which give the best prediction results are then saved, combined and used in an Ensemble to further push prediction performance.

Data Preparation

The dataset was provided by "Empirica ag", which is a corporation for research and consulting in Germany. The dataset contains data from the years 2016–2019, 2.366.122 samples and 23 features. According to Empirica, the sources of the data are auditors, real estate and housing companies, researchers, brokerage houses, banks and federal communes [Empirica ag, 2019]. The dataset is supposed to be curated and maintained by statisticians as well as checked for consistency to guarantee its integrity.

The features contain properties like sale price, living space, number of rooms, garden, construction year, second bathroom, etc. A full list of all features with descriptions can be found in the appendix. In Fig. 3 it is shown that numerous features contain missing values.

Figure 3. Feature display and missing values

```
RAW Shape:  (2366122, 23)
NaN Values:
 Postcode                  0
State                      0
County                  1092
City                     338
PropertyType               0
SalePrice                  0
LivingSpace                0
PriceSqm                   0
Rooms                  24734
Garden                     0
Floors               1218629
YearBuilt             270819
NewBuilding                0
NeedRenovation             0
Renovated                  0
SecondBathroom             0
BathroomWindow             0
Chimney                    0
YearModernization    1893914
Balcony                    0
Parquet                    0
MarbleFloor                0
YearSurvey                 0
dtype: int64
```

Source: Bedorf [2021].

While missing values in "County" can be filled in with postcode data from other samples the most prominent features in this regard are Rooms, Floors, „YearBuilt" and

„YearModernization“. The modernization year is set to the building year and vice versa to compensate for some of the missing values, while the feature „Floors“ needs to be dropped. After that all NaN values are removed.

For feature engineering the income data from each individual county is matched with the empirica dataset according to the survey year. Geographic coordinates are transformed into the metric “epsg:25832” standard format and matched with the postcodes. The area of each individual postcode area is then calculated and added as feature.

The distance of each postcode area to the nearest metropolitan area is calculated, as a measure of rurality.

Further additions are residents per postcode, building age and age demographics. This results in a data frame of 1.544.518 samples and 35 features. All demographic and financial data was provided by the Federal Statistical Office and can be found here [Federal Statistical Office Germany, 2021]. Geographic map data for distance calculations, postal area shapes was contributed by OpenStreetMap [OpenStreetMap, 2017].

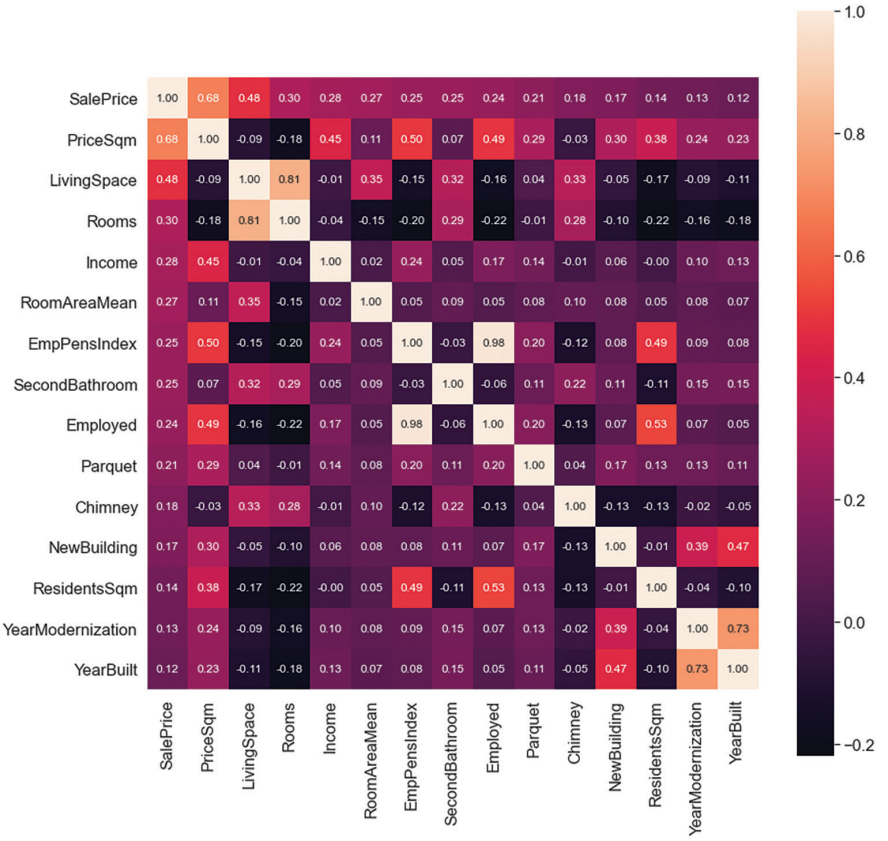
Data Analysis

To analyze the feature set we first compute the Pearson correlation coefficient [Laerd Statistics, 2021], which in this case is to be preferred over Kendall’s Tau [Magiya, 2019] due to the vast amount of samples in the dataset. Fig. 4 shows a heatmap of the first 15 features with the highest correlation.

We can see that “Income” correlates with “SalePrice” and “PriceSqm”, which indicates that the price of real estate is directly related to the available income per household in that region. “Employed” also has a high impact on PriceSqm. “Employed” is the population share between 20 and 50 years of age, thus the target group, which is more likely to acquire real estate properties.

Interestingly “YearBuilt”, the construction year does not seem to be a major factor which determines the sale price and is outnumbered by other minor features like Chimney or Parquet. There are also some cross correlations like SalePrice and PriceSqm, Rooms and LivingSpace, which must be taken care of.

Figure 4. Correlation Heatmap of 15 most important features



Source: Bedorf [2021].

Fig. 5 illustrates the correlation of the target variable. Focussing on the negative correlations one can see that the population share of pensioners negatively affects the "SalePrice", which makes intuitively sense, since a high share of pensioners would result in less demand on the market and thus lesser prices. There is also a negative correlation between the area, distance to metropolitan areas, building age and the geospatial y-coordinate. The area is calculated from the shape of individual postcodes and since cities in Germany are often divided into smaller postcodes this observation can be considered right as far as this empirical observation allows it. Building age and a high distance from city centers negatively affect the price. This supports common economic knowledge about the real estate market. The geospatial coordinate reflects the topological and economic conditions in Germany, where the south in general is wealthier than the north [The Economist, 2017].

Figure 5. Table of correlation to target

	SalePrice
SalePrice	1.000000
PriceSqm	0.677747
LivingSpace	0.476929
Rooms	0.300568
Income	0.283691
RoomAreaMean	0.274695
EmpPensIndex	0.251401
SecondBathroom	0.248989
Employed	0.237300
Parquet	0.213499
Chimney	0.178318
NewBuilding	0.171436
ResidentsSqm	0.139968
YearModernization	0.129918
YearBuilt	0.117601
Balcony	0.097798
MarbleFloor	0.091532
GeoX	0.080415
YearSurvey	0.050115
BathroomWindow	0.019303
Garden	0.013262
Residents	-0.001333
Renovated	-0.004872
NeedRenovation	-0.048831
GeoY	-0.103869
BuildingAge	-0.116594
DistanceNextCity	-0.130975
Area	-0.144327
Pensioner	-0.255513

Source: Bedorf [2021].

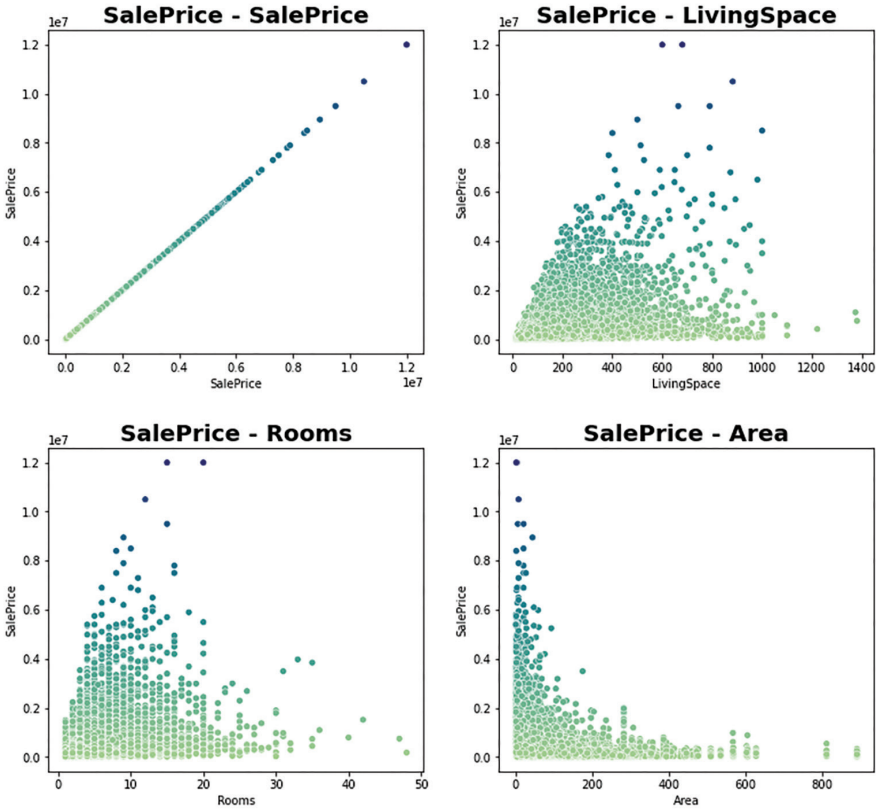
For outlier detection we use a visual approach by analyzing the contents of Fig. 6 and various boxplots. We furthermore transform the target to a normal distribution and cut off outliers which are not in a $\pm 2.7\sigma$ interval. Fig. 7 shows the target before and after the transformation and reduces the heavily skewed distribution from 7.9 to –0.22 and the kurtosis from 140 to 0.92.

Fig. 6 shows outliers which will negatively influence the prediction for some of the most important features. This incorporates real estates which have more than 700 m^2 of living space (0.1%), houses which are built before 1600 (0.07%) or have more than 20 rooms (0.06%). In fact, less than 1% of all samples have more than 8 rooms. This can also be interpreted as evidence for the high variance throughout the whole data set.

The sale price itself is also heavily influenced by outliers. The majority of all samples with more than 95% is below the border of 2 million euros, while less than 1000 samples cost more than 10 million euros.

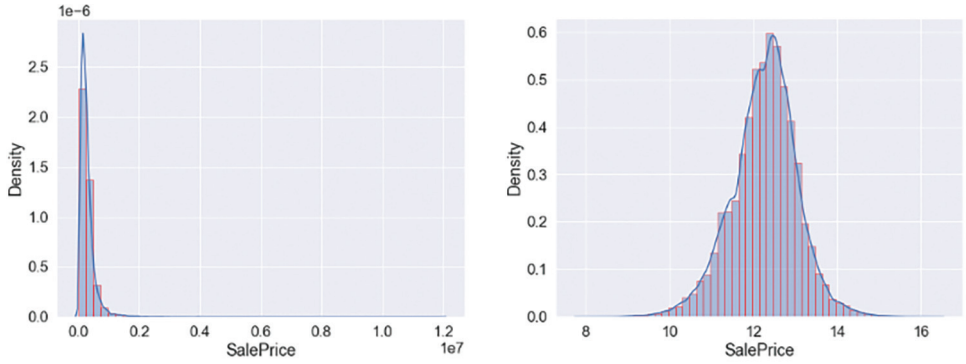
After outlier removal more than 98% (~1.500.000 samples) of the dataset is still available for prediction purposes. The high variance within many numeric features and the presence of outliers as shown in Fig. 8 emphasizes the need of this measure and further transformations like logarithmic functions.

Figure 6. Distribution of features



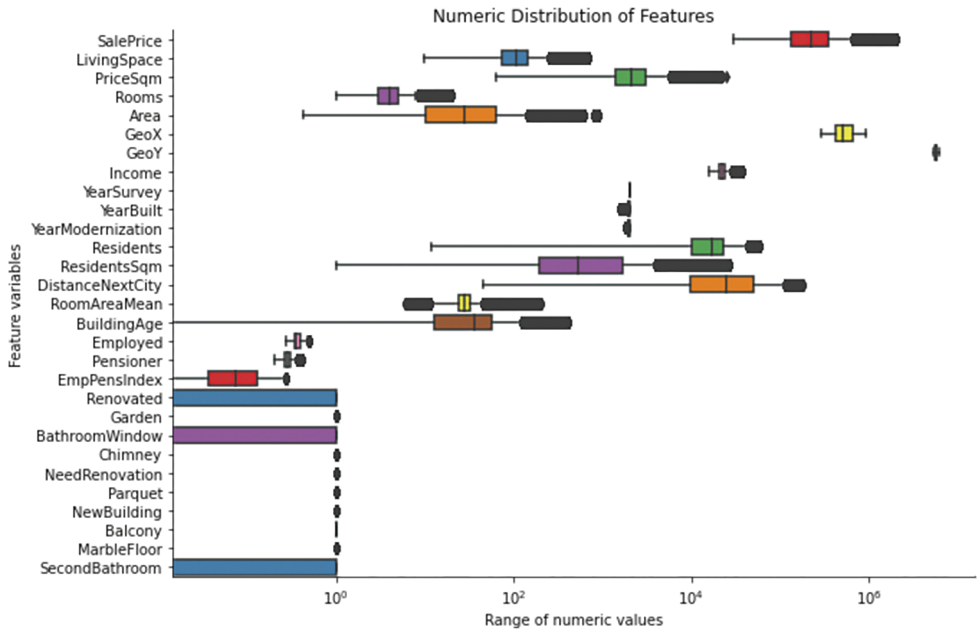
Source: Bedorf [2021].

Figure 7. SalePrice before and after log (x+1) transformation



Source: Bedorf [2021].

Figure 8. Logarithmic boxplot of numeric features



Source: Bedorf [2021].

Experiment

The experiments were performed on a cluster which was provided by the University of Applied Science Cologne. To conduct our experiments, we used the scikit-learn framework [Pedregosa, Varoquaux et al., 2011] and python 3.8. Results are directly visualized in a jupyter notebook for ease of use.

A pipeline of transformations [Shashanka, 2019] is used before the data is fit on the algorithms. These include Binary Encoding [Seger, 2018] for the categorical variable Postcode and One-Hot Encoding [Fawcett, 2021] for the "PropertyType". Target Encoding [Pargent, 2021] is used to fuse information about the price per square meter, based on individual postcodes into the training set. All features with a high, positive skew are log transformed and all features with a negative skew < –0.6 are power transformed via "yeo-johnson". In the last step all features are scaled with a robust approach by removing the median and are scaled according to quantile range.

The cross-validated grid search over all parameters is performed by using successive halving [Li, 2018]. Over several iterations parameter candidates are trained on a smaller part of the dataset. With each iteration the size of candidates is halved, while the size

of the dataset is doubled. This continues until one candidate remains, which is then trained on the whole dataset. This method makes the heavy workload grid searches introduce more feasible.

The algorithms which are used include Ridge Regression [Géron, 2019], Linear State Vector Machines [Géron, 2019], RANSAC [Choi, 2009], Bayesian Ridge Regression [Géron, 2019] and Lasso [Géron, 2019]. In total more than 250 different parameter combinations are examined to determine the best performing algorithm.

Because of the large sample size of the dataset training simple regression models is the only feasible solution in terms of time complexity. However, by doing this, we exclude our research from the performance gains, which are to be expected from modern algorithms like kernel based SVRs, Decision Tree Regressors or the popular XGBoost algorithm. All are expected to have at least quadratic time complexity and/or high memory requirements [RUser4512, 2018].

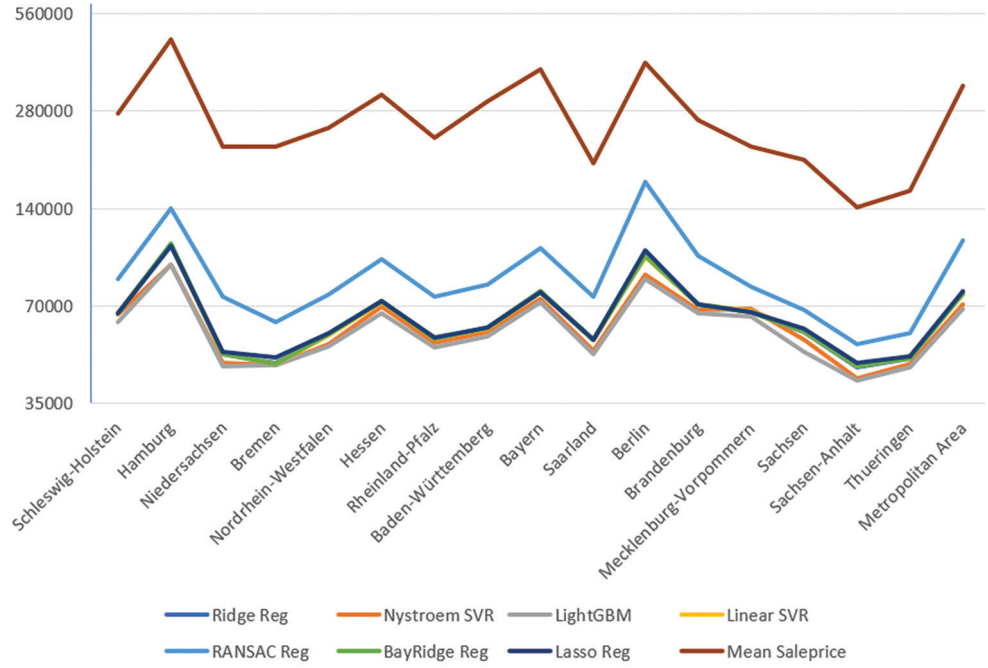
In order to approximate the common SVR the Nystroem method [Yang, Li et al., 2012] was used to create a low-rank rbf kernel approximation with 200 features. The method was modified to further improve performance as shown in this paper [He, Zhang, 2018], by replacing the Monte Carlo sampling with an KNN cluster algorithm to get the landmarks from the dataset, which provide the base vector for the approximation.

The last algorithm used is LightGBM [Ke, Meng et al., 2017], which was developed by Microsoft. LightGBM is Gradient Boosting algorithm which is based on decision trees, but focusses on performance and scales better with larger datasets. It also incorporates many advantages which XGBoost [Chen, Guestrin, 2016] has, like regularization, bagging, different loss metrics and sparse optimization.

Fig. 9 shows the prediction performance on the test set regarding all 16 federal states in Germany and the "metropolitan area" which is defined as samples within a 10 km radius around all cities with more than 200.000 residents. The mean absolute error can directly be interpreted as the difference to the sale price in euro and is plotted in comparison to the mean sale price (red line).

As to be expected the distribution of higher sale prices leads to higher deviations in the prediction. Higher prices in Berlin for example, result in errors which scale accordingly. The similarity in all curves emphasizes this observation. With the exception of the ridge regression algorithm all other proposed models perform on a comparable level, while the Nystroem SVR and LightGBM are slightly better with error margins between 41.000 € and 90.000 €. The plot also reveals much lower house prices in the structurally weaker east of Germany, namely Thueringen, Sachen-Anhalt and Sachsen.

Figure 9. Prediction performance in different regions with MAE



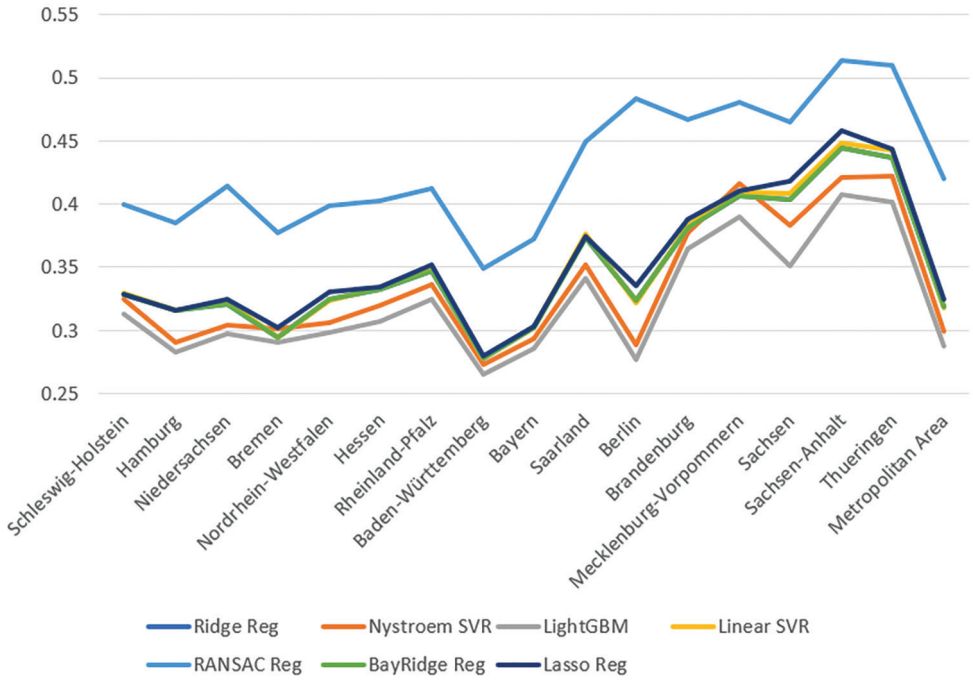
Source: Bedorf [2021].

Fig. 10 shows the same prediction, but this time with the RMSLE metric [Daniel, 2021], which is one of the most common used metrics for machine learning problems. RMSLE is useful for target variables with a wide range of values since it utilizes a log difference function to calculate the error. Big differences between target and prediction are thus penalized in the same way as small differences. In our case this is rather important since samples with high sales prices should not outweigh samples with small prices. RMSLE also penalizes underestimates more than over estimates. This is helpful since an overestimation of real estate should always be preferred to an under estimation.

As demonstrated in Fig. 9 Nystroem and LightGBM outperform all other algorithms. Because we are looking at percentages rather than absolute values the graphs now show a more reliable prediction quality. The worst performing states are located in the east of Germany. Reasons for that could be the more rural topography as well as the absence of samples in regions with less residents. The dataset clearly favors certain states in terms of sample size and quality.

A different training approach with sale prices between 100.000 € and 1.000.000 € only yielded slightly better results for LightGBM with a mean RMSLE score of 0.27.

Figure 10. Prediction performance in different regions with RMSLE



Source: Bedorf [2021].

Conclusion

The experiments have shown that modern machine learning algorithms outperform more established regression models by slight margins in exchange for computational overhead. The prediction of sale prices itself is not precise enough yet and can rather be interpreted as an estimation or trend.

The reasons for this are more likely found in the data itself than in the applied machine learnings methods. This thought is supported by the introduction of socioeconomic features like Income per household, rate of pensioners, etc. The correlation of features with the sale price, which are not primarily related to the real estate property itself show that not all factors which determine the sale price are present in the data.

The challenge and the goal for future researchers will be to create a sufficient, large scale dataset that also includes data like income, exact geospatial coordinates, transport links, schools, parks and even more detailed information about the property like a quality index.

A differentiation between flats and houses, as well as other regional parts is also thinkable as a future work, which could also incorporate ensemble methods by combining algorithms which were mentioned in this paper.

References

Bedorf N. [2021], *XAI–Modellagnostische Verfahren zur Erklärbarkeit von Machine Learning Algorithmen*, mimeo.

Bundesamt für Justiz [2019], *Federal Data Protection Act*, https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html (accessed: 9.10.2021).

Chen T., Guestrin C. [2016], *XGBoost: A Scalable Tree Boosting System*, DOI: 10.1145/2939672.2939785.

Choi S. [2009], *Performance evaluation of RANSAC family*, British Machine Vision Conference, BMVC, London, UK, September 7–10, DOI: 10.5244/C.23.81.

Daniel S. [2021], *Difference between RMSE and RMSLE*, https://www.datascienceland.com/blog/difference-between-rmse-and-rmsle-656/ (accessed: 4.10.2021).

Empirica ag [2019], *General Information*, https://www.empirica-institut.de/thema/regionaldatenbank/datenbank-regionaldaten/ (accessed: 9.10.2021).

Fawcett A. [2021], *Data Science in 5 Minutes: What is One Hot Encoding?* https://www.educative.io/blog/one-hot-encoding (accessed: 22.09.2021).

Federal Statistical Office Germany [2021], *Public datasets*, https://www-genesis.destatis.de/genesis/online (accessed: 18.09.2021).

Géron A. [2019], *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition* – O'Reilly, ISBN: 9781492032649.

Gonzalez S., Garcia S., Del Ser J., Rokach L., Herrera F. [2020], *A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities*, DOI: https://doi.org/10.1016/j.inffus.2020.07.007.

He L., Zhang H. [2018], *Kernel K-Means Sampling for Nyström Approximation*, DOI:10.1109/TIP.2018.2796860.

Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T. [2017], *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, NIPS.

Laerd Statistics [2021], *Pearson correlation*, https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php (accessed: 12.09.2021).

Li L. [2018], CMU, *Massively Parallel Hyperparameter Optimization*, https://blog.ml.cmu.edu/2018/12/12/massively-parallel-hyperparameter-optimization/ (accessed: 2.10.2021).

Magiya J. [2019], *Kendal Rank Correlation Explained*, https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535 (accessed: 20.09.2021).

OpenStreetMap contributors [2017], *Planet dump*, https://www.planet.osm.org (accessed: 9.10.2021).

Pargent F. [2021], *Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features*, arXiv:2104.00629 [stat.ML].

Park B., Bae J.K. [2015], *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data*, DOI: https://doi.org/10.1016/j.eswa.2014.11.040.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. [2011], *Scikit-learn: Machine learning in Python*, "Journal of Machine Learning Research", no. 12 (Oct), pp. 2825–2830.

Pow N., Janulewicz E., Liu L. [2014], *Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal.*

RUser4512 [2018], *Computational complexity of machine learning algorithms*, https://www.thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms/ (accessed: 4.09.2021).

Seger C. [2018], KTH, EECS, *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*, OAI:DiVA.org:kth-237426.

Shashanka M. [2019], *What is a Pipeline in Machine Learning? How to create one?* https://medium.com/analytics-vidhya/what-is-a-pipeline-in-machine-learning-how-to-create-one-bda91d0ceaca (accessed: 14.09.2021).

The Economist [2017], *On almost every indicator, Germany's south is doing better than its north*, https://www.economist.com/kaffeeklatsch/2017/08/20/on-almost-every-indicator-germanys-south-is-doing-better-than-its-north (accessed: 25.09.2021).

Viktorovich P.A., Aleksandrovich P.V., Leopoldovich K.I., Vasilevna P.I. [2018], *Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning*, DOI: 10.1109/RPC.2018.8482191.

Wu J.Y. [2017], *Housing Price prediction Using Support Vector Regression*, DOI: https://doi.org/10.31979/etd.vpub-6bgs.

Yang T., Li Y., Mahdavi M., Jin R., Zhou Z. [2012], *Nystroem Method vs Random Fourier Features: A Theoretical and Empirical Comparison, Advances in Neural Information Processing Systems*.

Yeo I.K., Johnson R.A. [2000], *A new family of power transformations to improve normality or symmetry*, "Biometrika", vol. 87(4), pp. 954–959.

Zhang C., Ma Y. [2012], *Ensemble Machine Learning, Methods and Applications*, Springer, ISBN: 978-1-4419-9326-7.

Appendix

Feature Table

Feature	Description
Postcode	Postcode with 5 digits
Postcode_3	Federal postcode, but only first three digits
State	One of 16 states in Germany.
County	District or precinct
City	Local community or city

Feature	Description
PropertyType	Townhouse, Flat, Detached House, etc.
SalePrice	Sale Price of Real Estate in Euro.
LivingSpace	The living space in $[m^2]$.
PriceSqm	$\frac{SalePrice}{LivingSpace}\left[\frac{€}{m^2}\right]$
Rooms	Number of rooms.
Area	Postal code area in $[km^2]$.
GeoX	X coordinate in EPSG:25832 standard according to ETRS89 in $[m^2]$.
GeoY	Y coordinate in EPSG:25832 standard according to ETRS89 in $[m^2]$.
Income	Available annual income per household in [€].
YearSurvey	Year of data collection.
YearBuilt	Construction year.
YearModernization	Modernization year.
Residents	Number of residents per postcode area.
ResidentsSqm	$\frac{Residents}{Area}\left[\frac{People}{km^2}\right]$
DistanceNextCity	Distance to next metropolitan area in [m].
RoomAreaMean	$\frac{Livingspace}{Rooms}$
BuildingAge	*YearSurvey – YearBuilt*.
Employed	Proportion of population that is between 20 and 50 years old.
Pensioner	Proportion of population that is older than 60 years.
EmpPensIndex	*Employed – Pensoner*.
Renovated	Is the building renovated?
Garden	Is there a garden?
BathroomWindow	Is there a windows in the bathroom?
Chimney	Is there a chimney?
NeedRenovation	Does the building need a renovation?
Parquet	Is there parquet floor?
NewBuilding	Is it a new building?
Balcony	Is there a balcony?
MarbleFloor	Is there marble floor?
SecondBathroom	Is there a second bathroom?