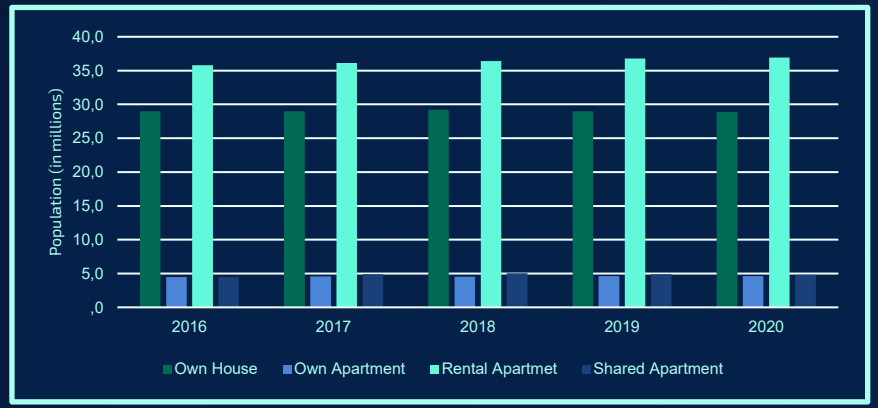
TABLE OF CONTENTS

01
INTRODUCTION

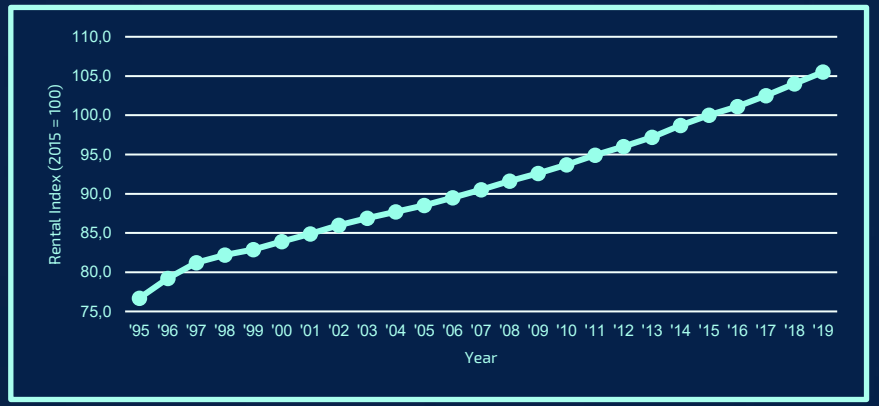
German Residential Rental Market

Germany is not a homeowner's market
but
instead a long term rental market
(Hachelberg et al., 2019, p. 17)

- Flexibility of the labour market
- Changing family structures
- Young people prefer to rent rather than own a flat
- Tenant - friendly laws in Germany
- In 2020, approximately 55% of the population lives in shared and rental apartments



Population in Germany by housing situation from 2016 to 2020 (IfD Al-lensbach, 2020)



Residential rental index for Germany in the years from 1995 to 2019 (Statistisches Bundesamt, 2020)

House Listing Websites

- Easy for both by tenants and landloards
- Easy access to rental units throughout Germany

REAL ESTATE BIG DATA

- Centralized
- Digitalized, free or machine-readable
- Updated in real-time, is high in volume
- Rich geo-location information

Cajias (2019)

To analyse large datasets, robust methodologies and tools are required which are far more complex than traditional hedonic regression models usage of machine learning techniques to analyse the large housing datasets

Objectives

Machine Leaning for RE Big Data Analysis

Comparison of Different Regression Methodologies for Rental Price Prediction in the German Residential Market

Attribute Impact

Parametric vs Nonparametric

02

BACKGROUND

Hedonic Theory

Lancaster (1966)

Consumers derive utility from the intrinsic attributes of the products

Rosen (1974)

Differentiated products comprise of various characteristics.

These characteristics, even though are not traded on markets, have an implicit marginal price attached to them that is revealed using hedonic regression.

The hedonic equation is determined by the bid of the consumers for different characteristics and the offers of these characteristics by suppliers.

Assuming an equilibrium market, maximization of utility results in the marginal bid by the consumers being equated to the marginal price by the sellers. Whereas, maximization of profits results in the marginal offer being equated to the marginal price.

Solving this maximization equation, gives the marginal price coefficients of each implicit characteristic.

Cropper et.al (1988, p. 669)

Empirical ambiguity:

- Form of utility function.
- Distribution of parameters of the utility function.
- Attributes considered for the analysis.
- Distribution of these attributes.
- Distribution of buyer characteristics

Form of utility function

Parametric

Models that require a finite set of parameters to generate a regression curve that is assumed to have a pre-existing functional form. The coefficients of the model are detemined by training the model on the dataset.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + e_i$$

Nonparametric

Models do not make assumptions regarding the form of the mapping function between input data and output. Consequently, they are free to learn any functional form from the training data.

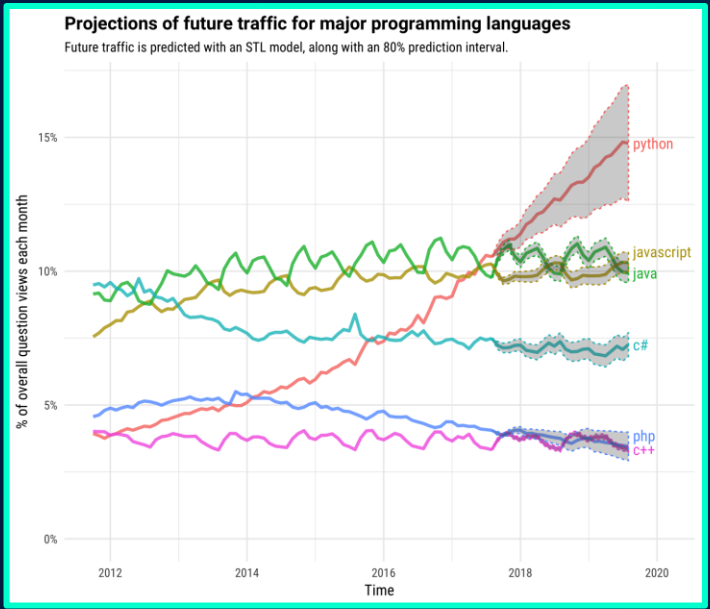
$$y = f(x_i) + e_i$$

Determined by Data

03
METHODOLOGY

Python

Python is a “interpreted, interactive, object-oriented programming language” with easy to use syntax (van Rossum & Drake, 2011). The language has been developed since two decades and is now become a robust integration platform which can be used to carry out numerous tasks like data gathering, data management and statistical analysis (Lin, 2012).



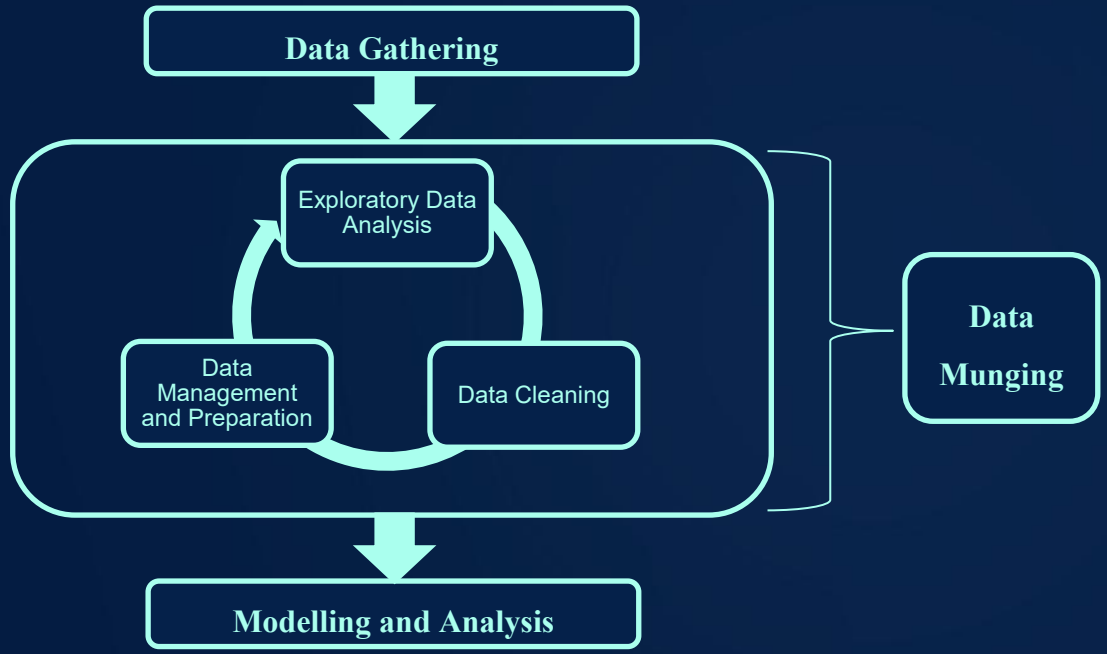
Source : (Stack Overflow, 2017)

Python Libraries

Library	Purpose
bs4	web scraping
geopandas	geospatial data operations
matplotlib	data visualization and plotting
numpy	scientific computing
pandas	data analysis and manipulation
scipy	scientific computing
seaborn	data visualization
scikit learn	regression analysis
xgboost	XGBoost regression

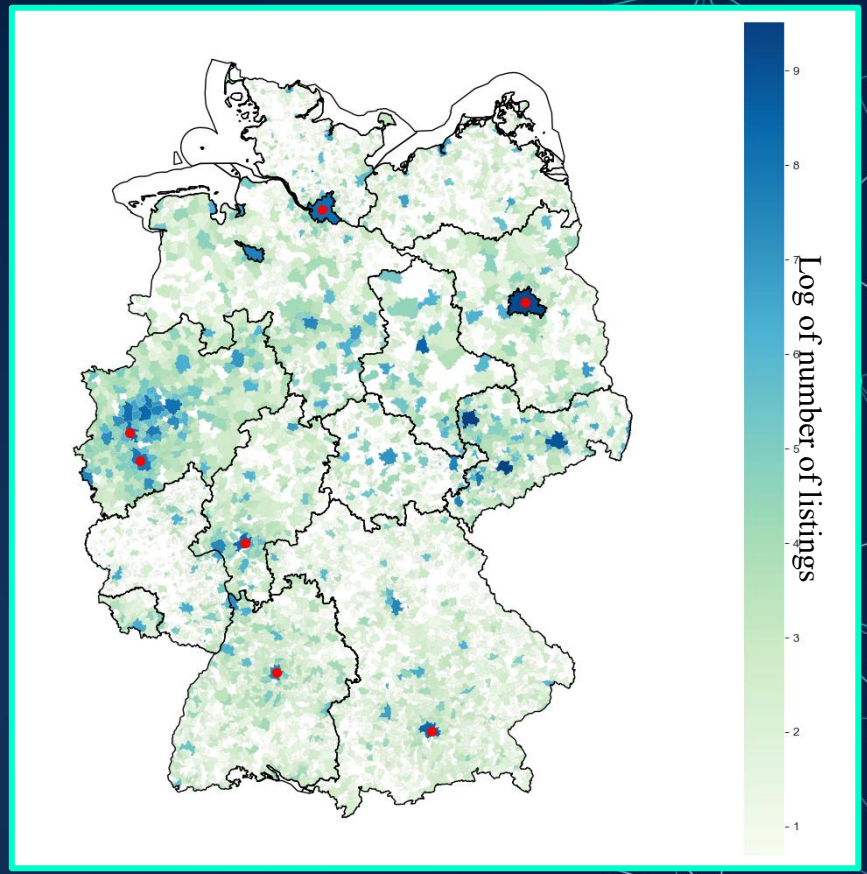
Analytical workflow

- Data munging is a systematic way of managing, preparing, exploring and cleaning the data before using it for analysis (Cady, 2017, pp. 47–60)



House Listing Data

- Immoscout24.com was used as the datasource
- Web scraping was performed in September 2018, May 2019, October 2019 and February 2020.
- A dataset of 268,850 rental apartments
- Each rental apartment had 48 features



Number of apartment listings per municipality

House Listing Data

Objekt-Nr.: 2f6becc8-1003-4b59-8732-d06b69d25f68 | Scout-ID: 123708800

Gepflegte 2-Raum-Wohnung mit Einbauküche in Frankfurt

60488 Frankfurt, Hausen
Die vollständige Adresse der Immobilie erhalten Sie vom Anbieter.
Auf Karte zeigen · Was kostet ein Umzug hierher?

580 €	2	54 m²
Kaltmiete	Zi.	Fläche

Einbauküche

Typ:	Etagenwohnung	Bonitätsauskunft:	SCHUFA-BonitätsCheck anfordern
Etage:	1 von 1	Zimmer:	2
Wohnfläche ca.:	54 m²	Badezimmer:	1
Bezugsfrei ab:	01.12.2020	Haustiere:	Nein

Kosten

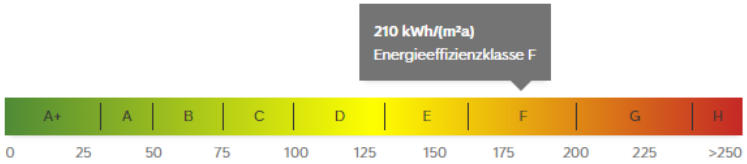
Kaltmiete:	580 €	Kaution o. Genossenschafts-anteile:	1740
Nebenkosten:	+ 150 €		Mieten ohne Kaution
Heizkosten:	in Nebenkosten enthalten	Umzugskosten:	Berechnung starten
Gesamtmiete:	**730 €**		

Was kostet Ihr Umzug?

VON Postleitzahl · NACH 60488 · **Preise vergleichen**

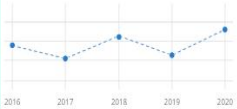
Bausubstanz & Energieausweis

Baujahr:	1960	Energieausweis:	liegt vor
Modernisierung/ Sanierung:	zuletzt 2015	Energieausweistyp:	Verbrauchsausweis
Objektzustand:	Gepflegt	Endenergie-verbrauch:	210 kWh/(m²*a)
Ausstattung:	Normale Qualität	Energieeffizienz-klasse:	F
Heizungsart:	Zentralheizung		
Wesentliche Energieträger:	Gas		



House Listing Data

Preistrend für Bestandsimmobilien in Hausen



› Preisentwicklung im Preisatlas ansehen

Sie benötigen Preisinformationen um zu entscheiden, ob Sie Ihre Immobilie verkaufen?

Wir beantworten alle Fragen zum Verkaufsprozess und unterstützen Sie Schritt für Schritt.

› Zur Verkäuferwelt

Karte

60488 Frankfurt, Hausen

› Mehr Information zu dieser Region

Die vollständige Adresse der Immobilie erhalten Sie vom Anbieter.

Internet Mobilfunk Preise/m² Wheelmap

Herr Gerhard Schäfer

Anbieter kontaktieren

Objektbeschreibung

Es handelt sich um eine 2-Zimmer-Wohnung (54qm) in der 1. Etage.

Die Wohnung besitzt eine Einbauküche, ein gefliestes Wannenbad (mit Duschwand) und mit Tageslicht (inkl. Waschmaschinenanschluss). Geheizt wird über eine Gaszentralheizung.

weiterlesen...

Ausstattung

Einbauküche

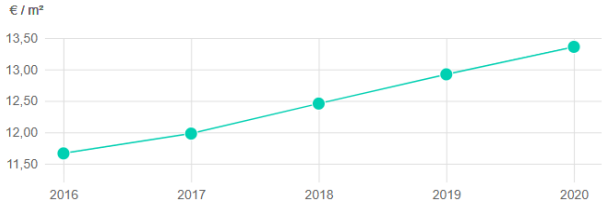
Lage

Das Wohnhaus liegt in U-Bahn- (Linie U6) und Bus-Nähe (Linien 72 und 73) und ist in Laufnähe zur Nidda und dem Niddapark, sowie dem Ortskern von Hausen mit Apotheke, Ärzten, Rewe und Aldi.

Entwicklung der Mietpreise für Wohnungen

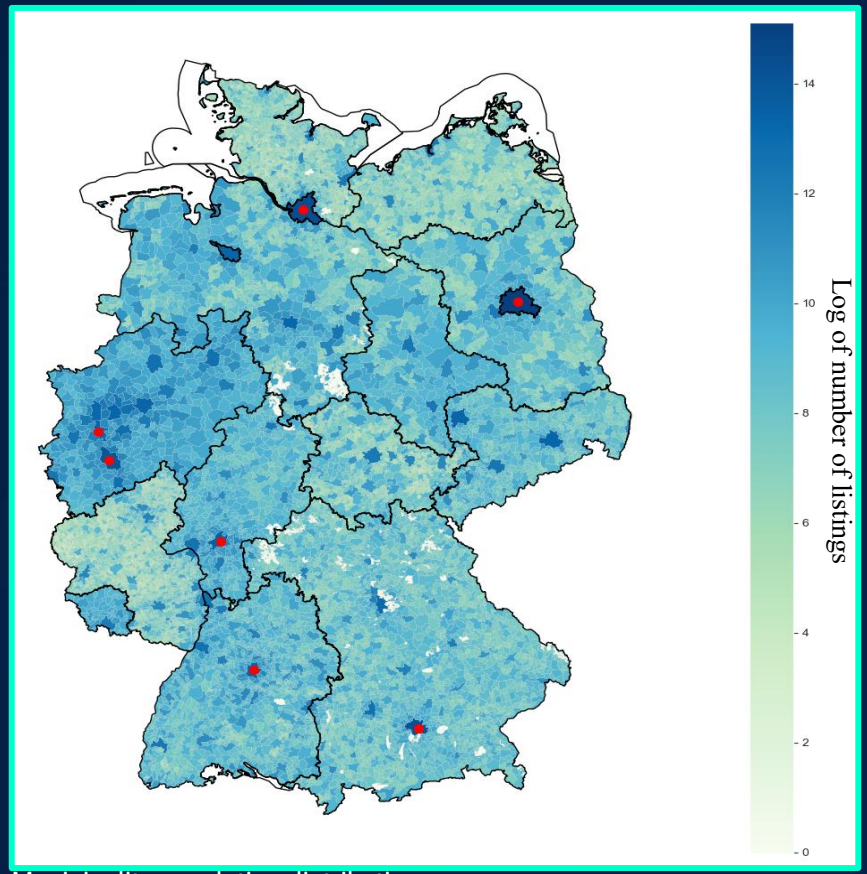
auf Basis durchschnittlicher, historischer Angebotspreise für Eigentumswohnungen in Hausen und Umgebung.

Q2 2020	zu Q2 2019
ø 13,42 € /m²	+5%



Spatial Data

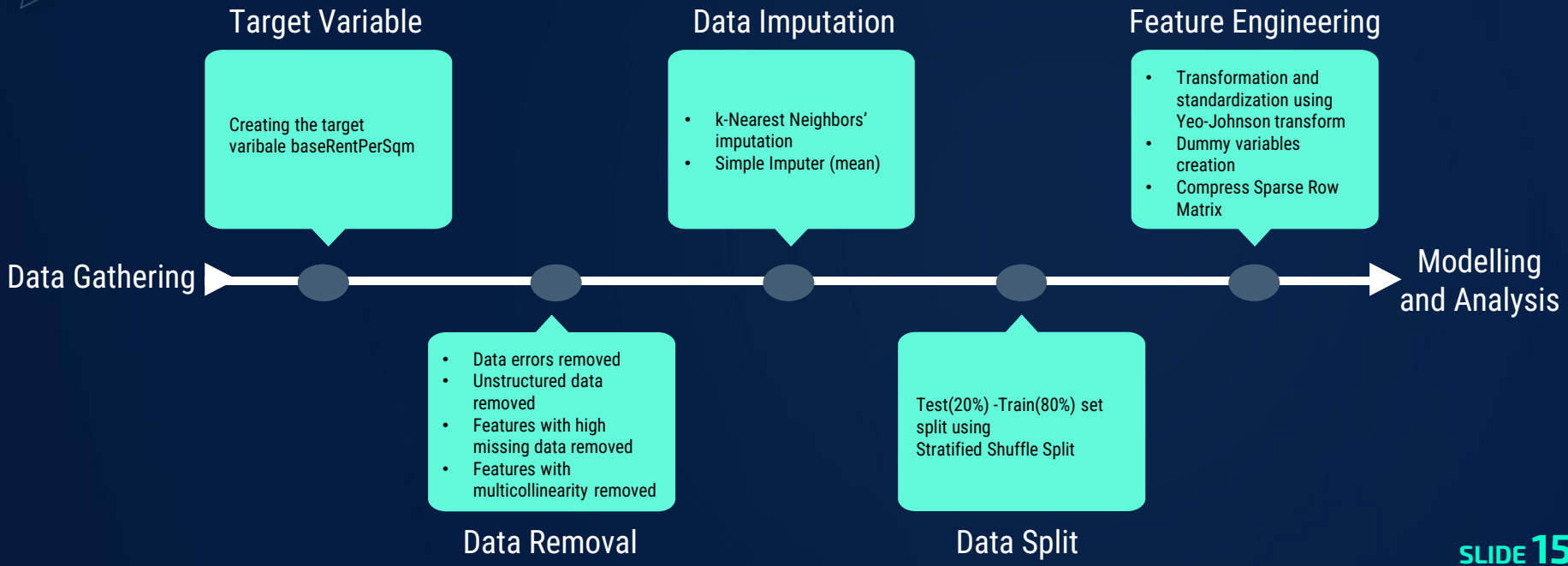
- A dataset called the *Verwaltungsgebiete mit Einwohnerzahlen*, which is published by Federal Agency for Cartography and Geodesy was used as the source of spatial data
- The dataset is in SHAPE file format consisting of geometry (polygons)



Municipality population distribution

Data Munging

Data Munging accounts for 50% to 80% of the time and cost in data analysis project and is very laborious (Kandel et al., 2011, p. 3363)



Final Dataset

246,333 rows x 225 columns and occupied 17.38 megabytes of memory

Numerical

Feature Name	count	mean	Std	min	25%	50%	75%	max
Target Feature								
baseRentPerSqm (€/m²)	246333	8.85	4.75	2.00	5.70	7.50	10.39	50.00
Independent Features								
floor	199276	2.09	1.65	-1.00	1.00	2.00	3.00	45.00
livingSpace	246333	73.79	32.06	10.00	54.08	67.97	87.00	542.00
noRooms	246333	2.62	0.98	1.00	2.00	3.00	3.00	10.00
numberOfFloors	156852	3.52	1.93	0.00	2.00	3.00	4.00	54.00
picturecount	246333	9.86	6.45	0.00	6.00	9.00	13.00	121.00
pop_density	246333	1409.91	1106.02	9.45	508.10	1118.47	2016.54	4736.11
pricetrend	244660	3.44	1.98	-12.33	2.05	3.44	4.62	14.92
serviceCharge (€)	246333	151.34	88.58	0.00	96.00	137.00	190.00	3500.00
thermalChar (kWh/(m² a)	246333	114.35	51.85	0.10	85.00	115.70	135.44	1983.00
yearConstructed	194257	1967	42.25	1500	1950	1972	1997	2090
telekomUploadSpeed	246333	25.67	17.87	0	2.4	40	40	100

Categorical

Feature Name	count	unique	top	frequency
balcony	246333	2	True	153019
cellar	246333	2	True	158807
condition	246333	11	NO INFORMATION	63140
date	246333	4	Feb20	72242
energyEfficiencyClass	246333	9	D	62760
firingTypes	246333	128	gas	102059
garden	246333	2	False	197545
geo_bln	246333	16	Nordrhein_Westfalen	60686
hasKitchen	246333	2	False	160326
heatingType	246333	14	central_heating	118202
interiorQual	246333	5	NO INFORMATION	103008
lift	246333	2	False	187047
newlyConst	246333	2	False	226467
petsAllowed	246333	4	NO INFORMATION	104640
typeOfFlat	246333	11	apartment	121376
telekomTvOffer	246333	3	ONE_YEAR_FREE	210330

04

REGRESSION MODELS

Machine Learning Algorithms

"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort" (Samuel, 1959, p. 8)

- Supervised Learning
- Unsupervised Learning
- Semi Supervised Learning
- Reinforced Learning
- Transduction
- Learning to learn

Parametric

Ordinary Least Square Linear

Stochastic Gradient Descent

Elastic Net SGD

Nonparametric

Linear Support Vector

Random Forest

XGBoost

Source : (Zhang, 2010, pp. 19–22)

Ordinary Least Square Linear

Functional Form

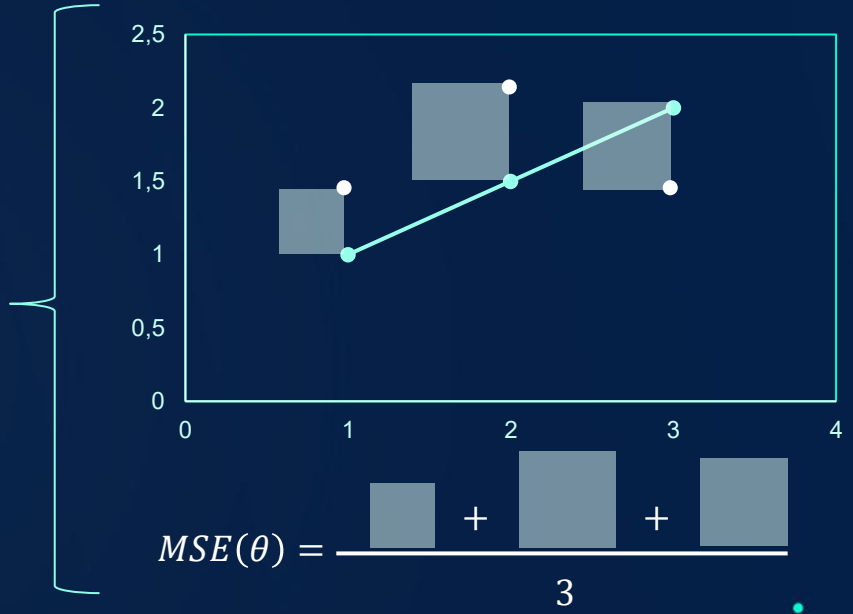
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Cost Function

$$MSE(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

Parameters

Parameter Name	Parameter Value
fit_intercept	True
normalize	False
copy_X	False
n_jobs	-1



Source : (Géron, 2019, pp. 167–168)

Stochastic Gradient Descent

Functional Form

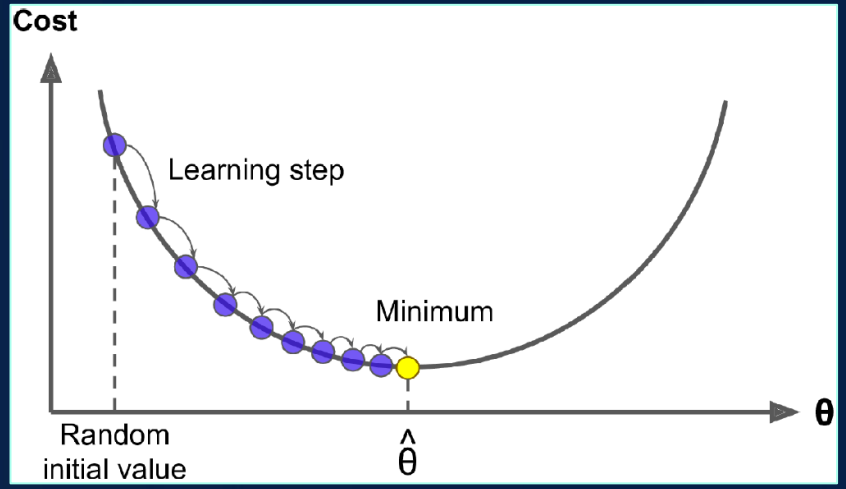
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Cost Function

$$MSE(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

Parameters

Parameter Name	Parameter Value
loss	squared_loss
eta0	0.001
tol	0.00001
max_iter	10000
n_jobs	-1
penalty	None



$$\theta(\text{next step}) = \theta - \eta \nabla\theta \, \text{MSE}(\theta)$$

Source : (Géron, 2019, pp. 173–188)

Elastic Net SGD

Functional Form

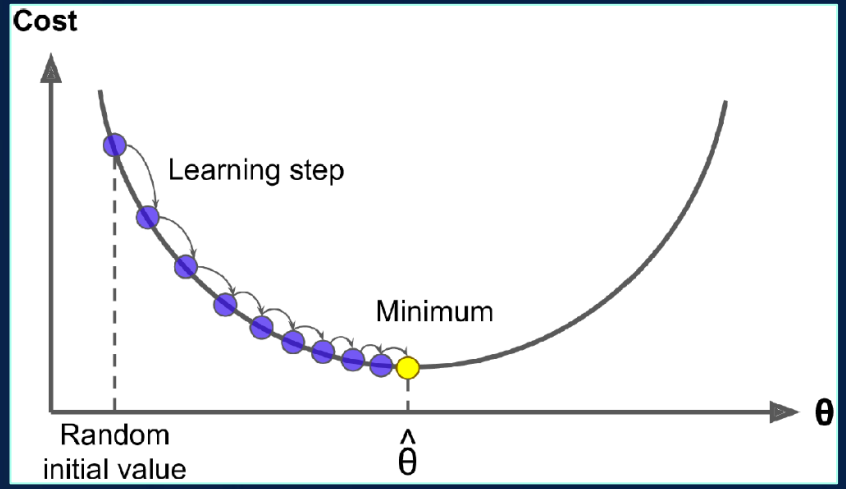
$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Cost Function

$$J(\theta) = MSE(\theta) + \underbrace{r\alpha \sum_{i=1}^{n} |\theta i|}_{L_1} + \underbrace{\alpha \frac{1-r}{2} \sum_{i=1}^{n} \theta i^2}_{L_2}$$

Parameters

Parameter Name	Parameter Value
loss	squared_loss
eta0	0.001
tol	0.00001
max_iter	10000
n_jobs	-1
penalty	None
alpha	0.0001
l1_ratio	0.1



$$\theta(\text{next step}) = \theta - \eta \nabla \theta \, \text{MSE}(\theta)$$

L_2 : Ridge algorithm

L_1 : Least Absolute Shrinkage and Selection Operator

Source : (Géron, 2019, pp. 199–202)

Linear Support Vector Regression

Functional Form

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Cost Function

Hard margin / Soft margin

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i)$$

$$\text{subject to} \quad y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i,$$

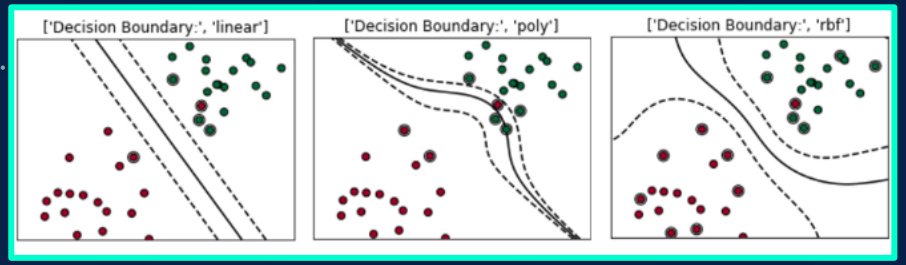
$$\zeta_i \geq 0, i = 1, \ldots, n$$

Parameters

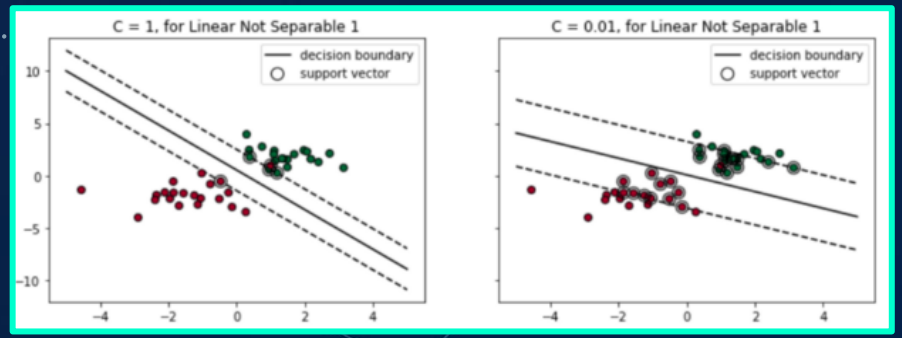
Parameter Name	Parameter Value
C	0.5
epsilon	0.5
loss	epsilon_insensitive
max_iter	10000

Source : (Smola & Schölkopf, 2004, pp. 199–222)

Kernel Trick



Soft Margin



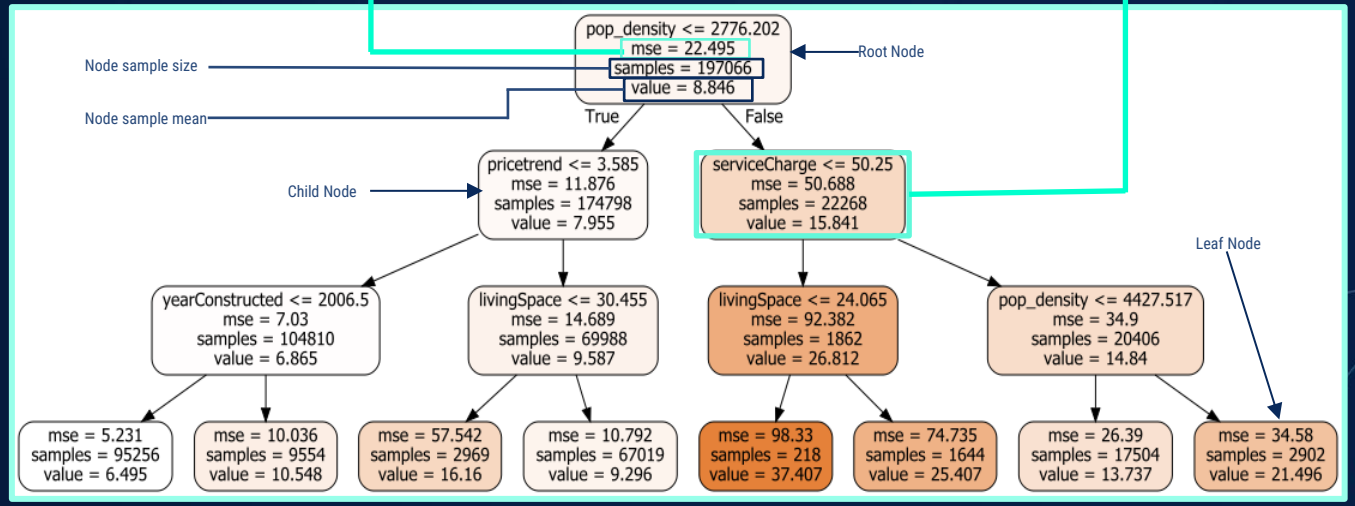
(Chen, 2019)

Random Forest Regression

DECISION TREE is the Building Block for Random Forest

Classification and Regression Tree (CART) algorithm

$$J = \frac{m_{left}}{m}\text{MSE}_{left} + \frac{m_{right}}{m}\text{MSE}_{right}$$



Decision Trees overfit the giving dataset and hence cannot generalize (Géron, 2019, p. 249).

Source : (Géron, 2019, pp. 240–251)

Random Forest Regression

Breiman (2001, p. 1):

Random forests are a combination (Ensemble) of tree (Decision Tree) predictors such that each tree depends on the values of random vector sampled independently and with the same distribution for all trees in the forest.

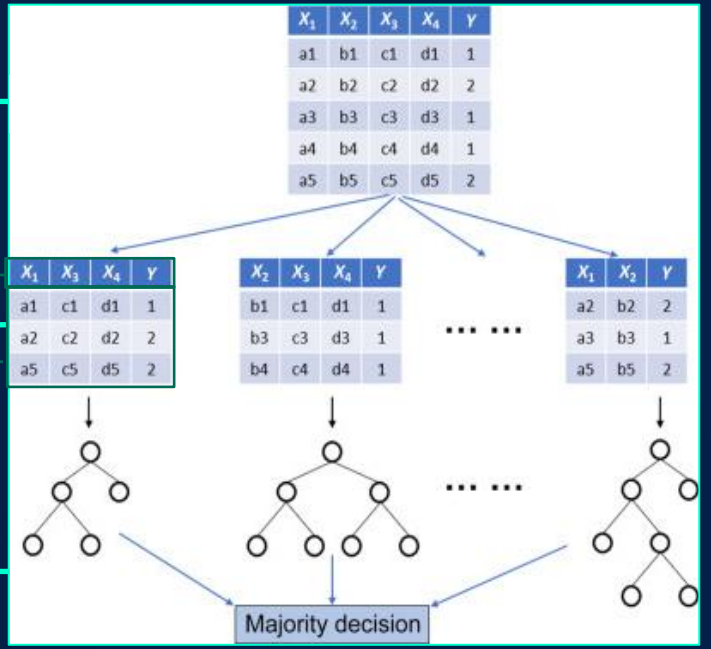
Complete Dataset

Bootstrap Aggregation

Random subset of features

Random subset of data

Ensemble of Trees



(Misra & Li, 2020, pp. 243–287)

Source : (Géron, 2019, pp. 240–251)

Random Forest Regression

Parameters

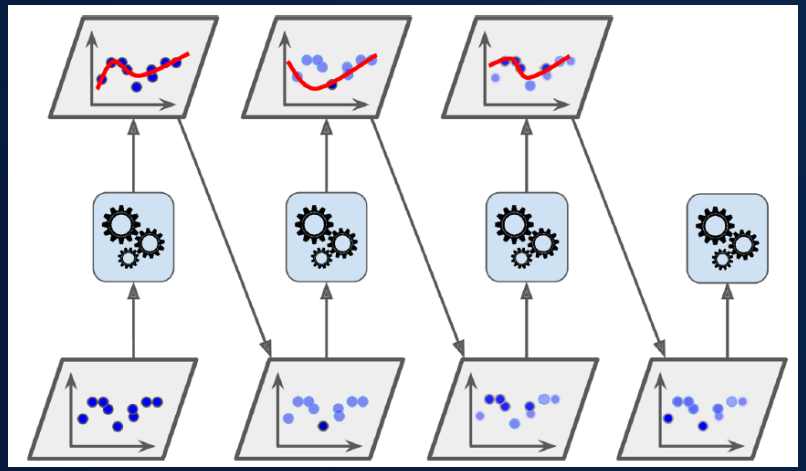
Parameter Name	Parameter Value
n_estimators	10
criterion	mse
max_features	sqrt
bootstrap	True
n_jobs	-1

Source : (Géron, 2019, pp. 240–251)

XGBoost Regresion

Developed by Chen & Guestrin (2016), **XGBoost** *(Extreme Gradient Boosting) is a scalable end to end* **Gradient Boosted Decision Trees (GBDT)** *algorithm widely used by data scientists*

Gradient Boosting *methods is to train predictors* **sequentially**, *each trying to correct its predecessor based on the residual errors (Géron, 2019, p. 263)*



(Géron, 2019, pp. 264)

Source : (Géron, 2019, pp. 263-264)

GBDT + optimizations = XGBoost

- An improvised tree learning algorithm for handling sparse data.
- Faster model exploration using parallel and distributed computing.
- Enabling processing of hundred million of examples by exploiting cache-aware block structure for out-of-core tree learning.

XGBoost Regresion

Parameters

Parameter Name	Parameter Value
n_estimators	100
colsample_bytree	0.5
gamma	0.0468
learning_rate	0.05
max_depth	20
nthreads	-1
random_state	7
reg_alpha	0.5
reg_lambda	0.5
silent	1
subsample	0.8

Source : (Géron, 2019, pp. 240–251)

Performance Measurement

Mean Square Error

$$\mathrm{MSE}\left(\mathrm{y},\hat{\mathrm{y}}\right) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Square Error

$$\mathrm{RMSM}\left(\mathrm{y},\hat{\mathrm{y}}\right) = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=0}^{\mathrm{n}-1}(y_i - \hat{y}_i)^2}$$

Coefficient of Determination

$$\mathrm{R}^2(\mathrm{y},\hat{\mathrm{y}}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

TRAINING

Iterative Process using

Grid Search Hyper-parameter Tuning

To reach Optimal Model

TESTING

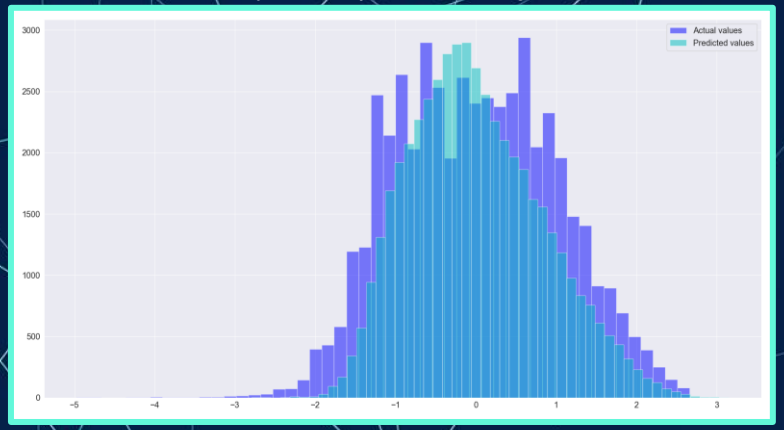
05
RESULTS & DISCUSSION

Performance

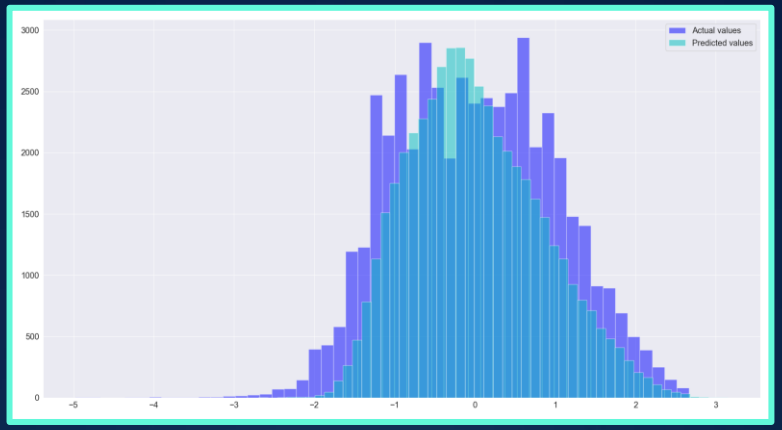
Type of Regression	Training Scores		Testing Scores		
	MSE	RMSE	MSE	RMSE	R^2
Ordinary Least Square Linear Regression	0.3080	0.5550	0.3133	0.5598	68.64%
Stochastic Gradient Descent Regression	0.3083	0.5553	0.3138	0.5602	68.59%
Elastic Net SGD Regression	0.3084	0.5553	0.3137	0.5601	68.61%
Linear Support Vector Regression	0.3082	0.5551	0.3137	0.5600	68.60%
Random Forest Regression	0.2222	0.4715	0.2147	0.4634	78.51%
XGBoost Regression	0.1477	0.3843	0.1440	0.3795	85.58%

Predicted Versus Actual

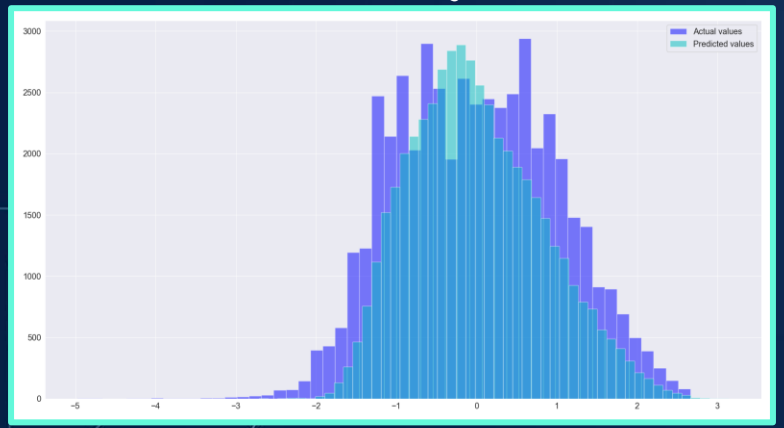
Ordinary Least Square Linear Regression



Stochastic Gradient Descent Regression

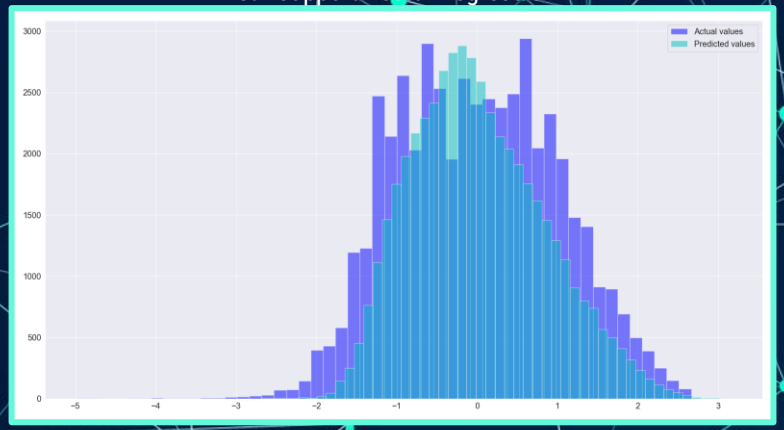


Elastic Net SGD Regression

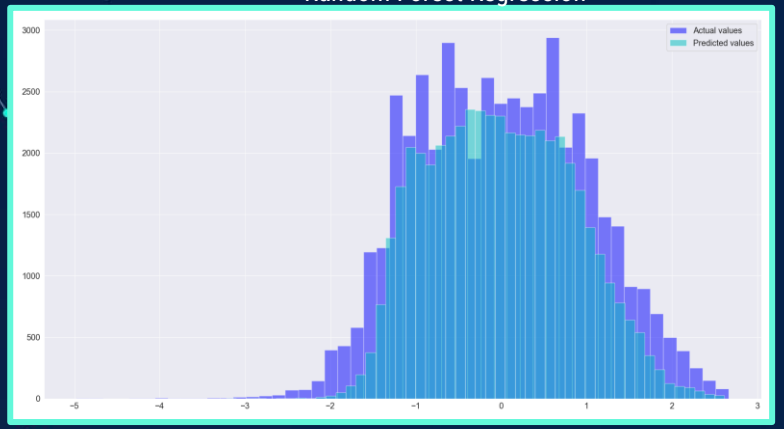


Predicted Versus Actual

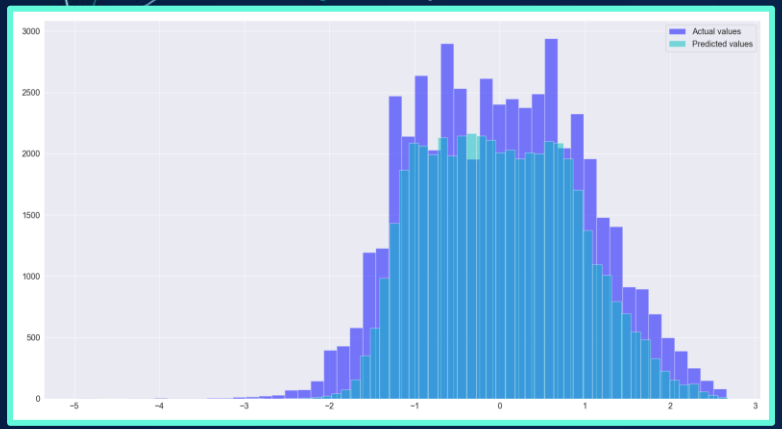
Linear Support Vector Regression



Random Forest Regression

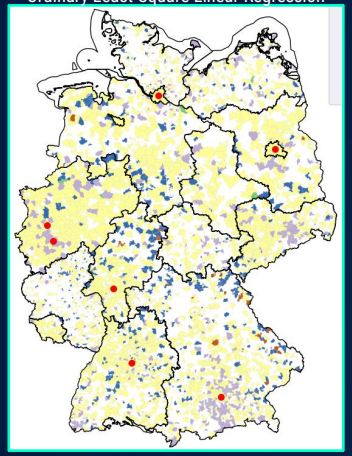


XGBoost Regression

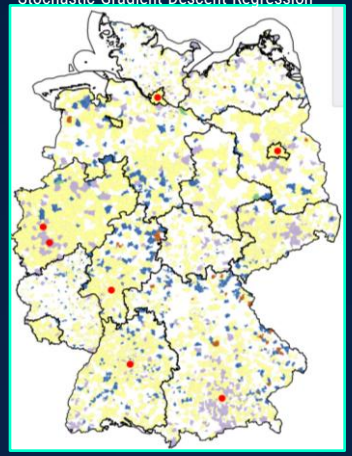


Spatial Representation

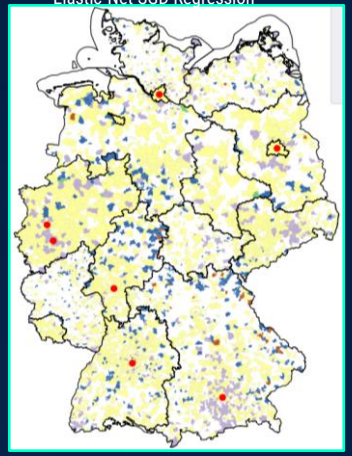
Ordinary Least Square Linear Regression



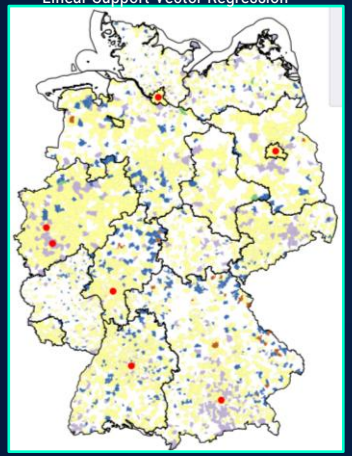
Stochastic Gradient Descent Regression



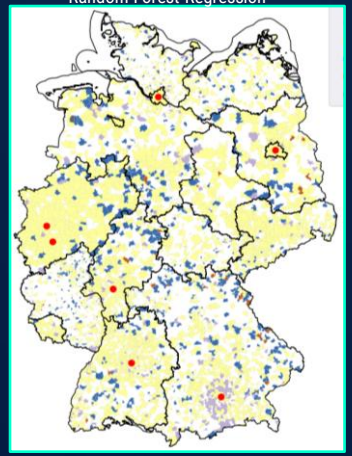
Elastic Net SGD Regression



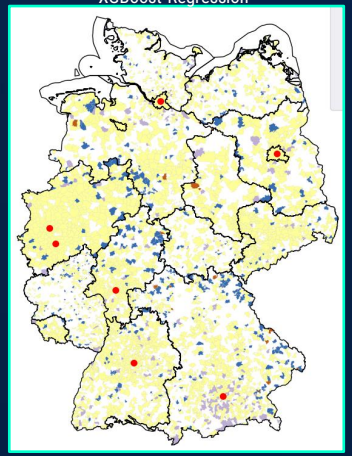
Linear Support Vector Regression



Random Forest Regression



XGBoost Regression

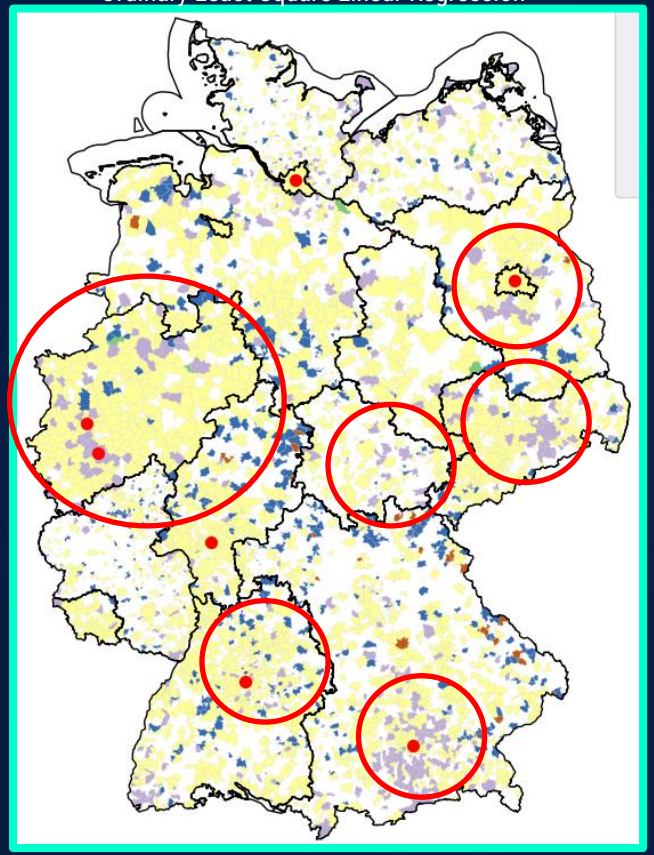


$$y_{diff} = \frac{\sum_{i=1}^{n} y_{pred} - y_{actual}}{n}$$

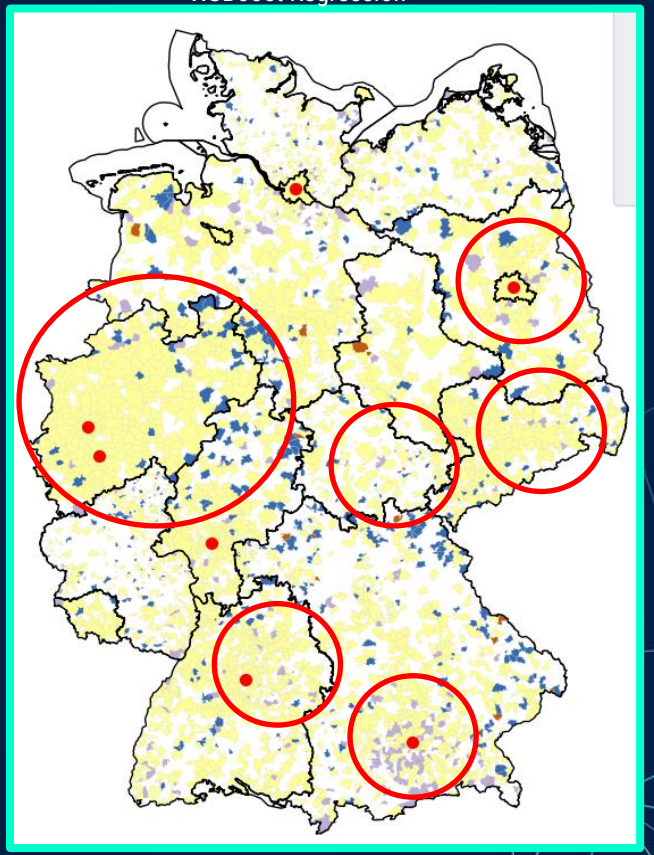
Color	y diff	Note
	Less than -2.5	Complete false prediction
	-2.5 to -1.5	predicted rents much lower than actual rents
	-1.5 to -0.5	predicted rents lower than actual rents
	-0.5 to 0.5	predicted rents almost equal to actual rents
	0.5 to 1.5	predicted rents higher than actaul rents
	1.5 to 2.5	predicted rents much higher than actual rents
	More than 2.5	Complete false prediction
	-	Municipality not included in Test Set

Spatial Representation

Ordinary Least Square Linear Regression



XGBoost Regression



Attribute Analysis

FEATURE IMPORTANCE
XBOOST

FEATURE COEFFICIENT
Elastic Net SGD

IMPACT + DIRECTION = ATTRIBUTE ANALYSIS

REAL ESTATE TRENDS

Attribute Analysis

States

Feature Name	Score	Coeff
Sachsen	13.51%	-0.4523
Bayern	4.31%	0.3805
Hamburg	3.87%	0.3450
Baden Württemberg	3.63%	0.4174
Hessen	3.09%	0.3846
Berlin	2.82%	-0.2143
Nordrhein Westfalen	2.03%	-0.2125
Sachsen Anhalt	1.98%	-0.2997
Thüringen	1.44%	-0.2785
Niedersachsen	1.06%	0.0489
Schleswig Holstein	0.77%	0.0276
Rheinland Pfalz	0.71%	0.0873
Bremen	0.45%	-0.1943
Mecklenburg Vorpommern	0.43%	-0.0754
Brandenburg	0.32%	0.0221
Saarland	0.30%	0.0013
Total	40.75%	

Attribute Analysis

Miscellaneous

Feature Name	Score	Coeff
Newly constructed	8.98%	0.1569
Poplation density	2.94%	0.3211
Price trend	2.14%	0.2358
Year of Construction	0.84%	0.0152
Living space	0.44%	-0.1454
Service charge	0.44%	0.0504
Number of rooms	0.36%	-0.0386
thermal Char	0.28%	-0.0012
Number of floors	0.19%	-0.0252
Picture Count	0.17%	0.0550
Floor	0.14%	-0.0223
Total	16.92%	

Attribute Analysis

Amenities

Feature Name	Score	Coeff
kitchen	12.31%	0.2497
lift	1.48%	0.1891
balcony	0.51%	0.1035
cellar	0.20%	-0.0830
garden	0.15%	0.0134
Total	14.65%	

Attribute Analysis

Condition

Feature Name	Score	Coeff
First time use	1.28%	0.2158
First time use after refurbishment	0.41%	0.1997
Well kept	0.35%	-0.0783
Mint condition	0.34%	0.1312
Need of renovation	0.23%	-0.2194
NO INFORMATION	0.20%	-0.0295
Refurbished	0.17%	-0.0085
Negotiable	0.16%	-0.1440
Fully renovated	0.16%	-0.0369
Modernized	0.12%	-0.0300
Ripe for demolition	0.08%	0.0000
Total	3.50%	

Attribute Analysis

Interior Quality

Feature Name	Score	Coeff
NO INFORMATION	0.36%	-0.0698
Simple	0.42%	-0.2044
Normal	0.79%	-0.1365
Luxury	1.32%	0.2842
Sophisticated	2.35%	0.1206
Total	5.24%	

Attribute Analysis

Type of Flat

Feature Name	Score	Coeff
NO INFORMATION	0.19%	-0.0243
Penthouse	0.17%	0.1204
Maisonette	0.14%	0.1218
Ground floor	0.13%	-0.0363
Roof storey	0.12%	-0.0474
Loft	0.12%	0.0018
Terraced flat	0.12%	0.0380
Apartment	0.12%	-0.0574
Other	0.12%	-0.0342
Half basement	0.11%	-0.0440
Raised ground floor	0.11%	-0.0401
Total	1.45%	

Attribute Analysis

Energy Efficiency Class

Feature Name	Score	Coeff
A +	0.31%	0.1105
B	0.21%	0.0847
A	0.17%	0.0646
E	0.17%	-0.0058
D	0.13%	-0.0622
H	0.12%	-0.0643
F	0.11%	-0.0376
G	0.11%	-0.0527
C	0.10%	-0.0389
Total	1.41%	

Attribute Analysis

Heating Type

Feature Name	Score	Coeff
Floor heating	0.36%	0.0990
NO INFORMATION	0.26%	0.0736
Stove heating	0.22%	-0.0472
District heating	0.20%	-0.0186
Central heating	0.19%	0.0008
Night storage heater	0.18%	-0.1312
Combined heat and power plant	0.15%	-0.0245
Oil heating	0.14%	0.0074
Electric heating	0.13%	-0.0275
Self contained central heating	0.12%	-0.0118
Gas heating	0.11%	0.0372
Heat pump	0.11%	0.0393
Wood pellet heating	0.10%	-0.0001
Solar heating	0.08%	-0.0002
Total	2.35%	

Attribute Analysis

Pets

Feature Name	Score	Coeff
No	0.65%	0.0981
Negotiable	0.23%	-0.0602
NO INFORMATION	0.15%	0.0180
Yes	0.15%	-0.0638
Total	1.18%	

Attribute Analysis

Data Extraction Date

Feature Name	Score	Coeff
Feb 20	0.20%	0.0654
Sep 18	0.16%	-0.0714
Oct19	0.13%	0.0323
May19	0.13%	-0.0343
Total	0.62%	

06
CONCLUSION

Conclusion

Non parametric models perform significantly better than the parametric models.

Amongst the parametric models, **Elastic Net SGD** (MSE = 0.31, RMSE = 0.56 and R^2 = 68.59%) showed a good balance between performance and computational efficiency, while also being able to generalize well to unknown data.

Using an **ensemble** of individual machine learning algorithms significantly improved prediction capabilities.

XGBoost, which is a Gradient Boosting algorithm consisting of an ensemble of Decision Trees, had the highest performance score (MSE = 0.14, RMSE = 0.38 and R^2 = 85.58%) while also being computationally efficient.

Amongst all the features, **spatial features** had the most impact on the listing rental price.

Limitations and Future Scope

The house listing data was extracted only for four time points (09.2018, 05.2019, 10.2019 and 02.2020). A regular extraction and collection of real-time data could improve model training and testing. Also, the data source could be extended to more than one listing websites.

The paper only used one additional feature (population density) apart from the house listing data. The dataset could be extended to incorporate multiple socio-economic attributes.

Only six regression models were compared. This list could be extended to give a more comprehensive comparison of various regression models. For example, Artificial Neural Network models (Selim, 2009) could be explored for hedonic analysis.

One of the biggest limitations of this paper was the computational power (RAM: 8 Gigabytes, Processor: Intel (R) Core (TM) i5-6300 CPU @ 2.40GHz, 2.50GHz). More complex analysis can be carried out with higher computational power. These include better missing data imputation techniques, ensemble methods with a greater number of individual predictors, stacking of models and better hyper tuning of model.

To incorporate spatial information, only the German States (as dummy variables), were included. However, with higher computational power, spatial lags (on municipality level) of the endogenous or exogenous features can be incorporated in the regression.


Thank you

REFERENCES

Cady, F. (2017). *The Data Science Handbook*. John Wiley & Sons.

Chen, L. (2019, January 7). *Support Vector Machine—Simply Explained.* Medium. https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496

Cropper, M. L., Deck, L. B., & McConnell, K. E. (1988). On the Choice of Funtional Form for Hedonic Price Functions. *The Review of Economics and Statistics*, *70*(4), 668–675. JSTOR. https://doi.org/10.2307/1935831

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

IfD Allensbach. (2020, September 28). *Population in Germany by housing situation from 2016 to 2020 (in millions of people)*. Statista. https://de-statista-com.library.myebs.de/statistik/daten/studie/171237/umfrage/wohnsituation-der-bevoelkerung/

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3363–3372. https://doi.org/10.1145/1978942.1979444

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157. https://doi.org/10.1086/259131

REFERENCES

Lin, J. W.-B. (2012). Why Python Is the Next Wave in Earth Sciences Computing. *Bulletin of the American Meteorological Society*, *93*(12), 1823–1824. https://doi.org/10.1175/BAMS-D-12-00148.1

Misra, S., & Li, H. (2020). Chapter 9—Noninvasive fracture characterization based on the classification of sonic wave travel times. In S. Misra, H. Li, & J. He (Eds.), *Machine Learning for Subsurface Characterization* (pp. 243–287). Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. https://doi.org/10.1086/260169

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Stack Overflow. (2017, September 6). *The Incredible Growth of Python | Stack Overflow*. Stack Overflow Blog. https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Statistisches Bundesamt. (2020, February). *Entwicklung des Wohnungsmietindex für Deutschland in den Jahren von 1995 bis 2019 (2015 = Index 100)*. Statistisches Bundesamt. www.destatis.de

van Rossum, G., & Drake, F. L. (2011). *The Python Language Reference Manual*. Network Theory Ltd.

Zhang, Y. (2010). *New Advances in Machine Learning*. BoD – Books on Demand..