Interpretierbare Vorhersage von Immobilienpreisen mittels Cluster- und Predict-Methoden

22.11.2024 Paul Gümmer

Agenda

1. Motivation
2. Methodik
    a. Datensatz und Datenaufbereitung
    b. Cluster Ansätze und Bewertung
3. Ergebnisse
    a. Cluster-Then-Predict Ansatz
    b. Cluster-Cluster-Then-Predict Ansatz
    c. Clusterspezifische Merkmale
4. Ausblick

Motivation und Zielsetzung

- Einfache Modelle: interpretierbar, aber begrenzte Komplexitätsabdeckung
- Komplexe Modelle: erfassen Komplexität, jedoch oft nicht interpretierbar

- Cluster-Then-Predict Ansatz bereits durch Azimlu et al. (2021) in *"House price prediction using clustering and genetic programming along with conducting a comparative study"*

🎯 Beste Clustering-Algorithmen für die Segmentierung von Online-Immobilienangeboten finden

🎯 Einfluss von Merkmalen (Lage, Größe, Alter) auf die Preisvorhersage in Segmenten untersuchen

Methodik

RED CAMPUS Files Datensatz (Allgemein)

Originale Immobiliendaten (Haus-, Wohnungsverkäufe und Wohnungsvermietungen)

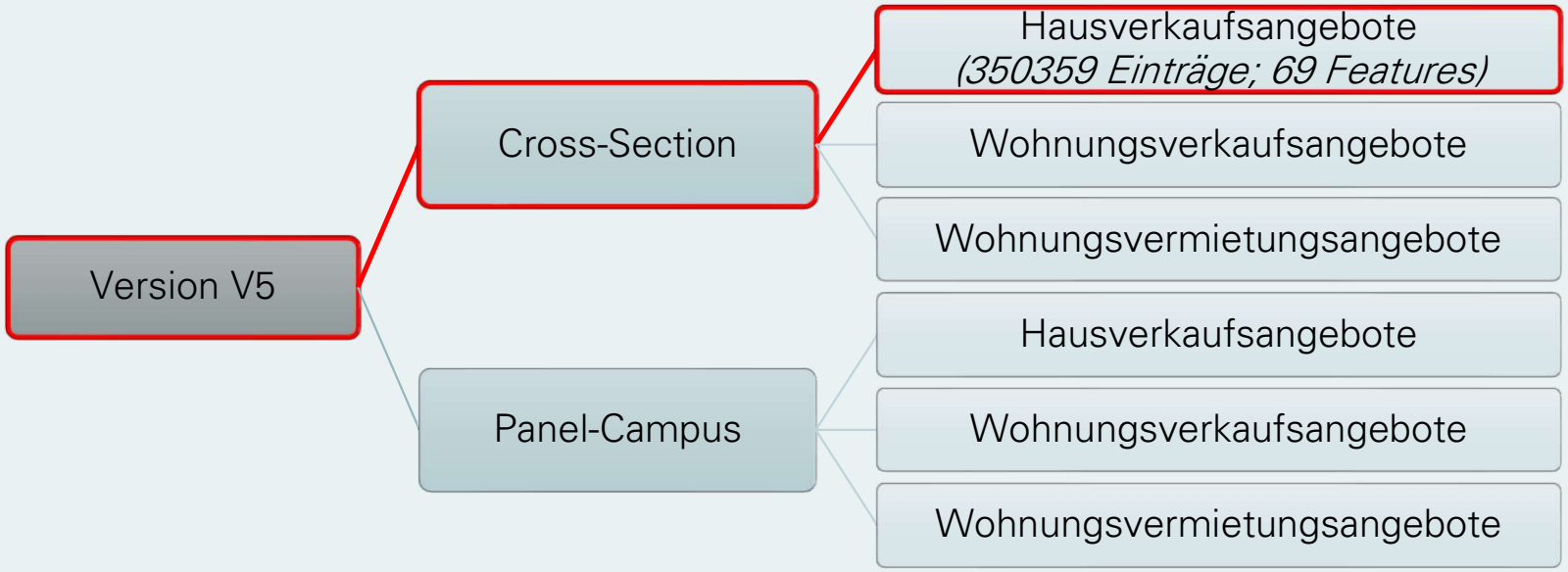
Daten stammen von ImmobilienScout24

Für nicht kommerzielle Zwecke nutzbar

Datenbeschaffung:

- Daten werden durch das RWI – Leibniz Institut für Wirtschaftsforschung bereitgestellt
- Zwei frei erhältliche Campus Files für Lehrende und Studierende

RED CAMPUS Files Datensatz (Struktur)



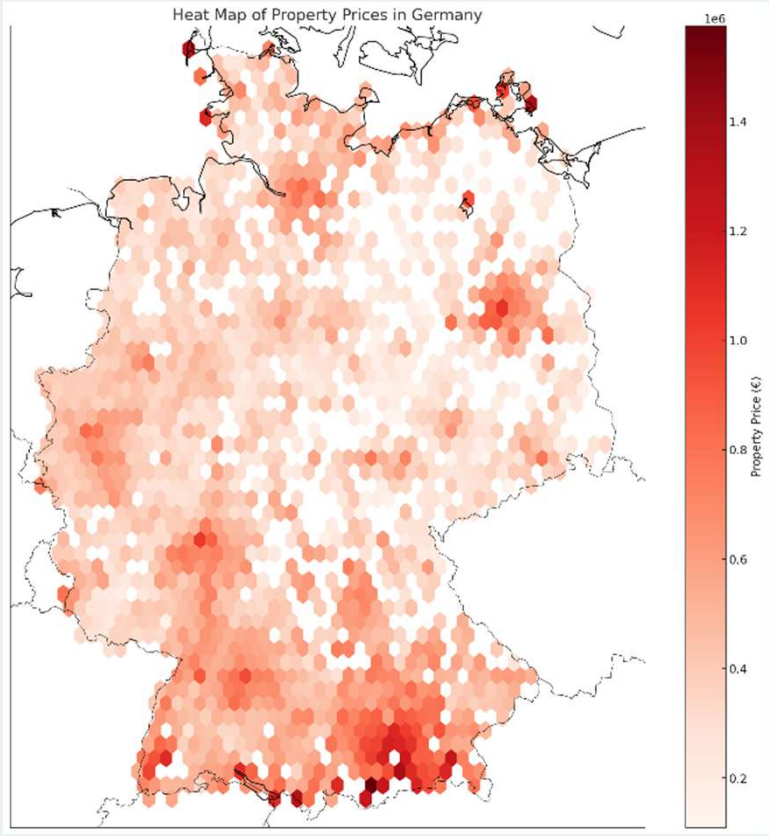
- Die Daten umfassen Preisinformationen, Geodaten sowie spezifische Charakteristika der Objekte

Datenaufbereitung

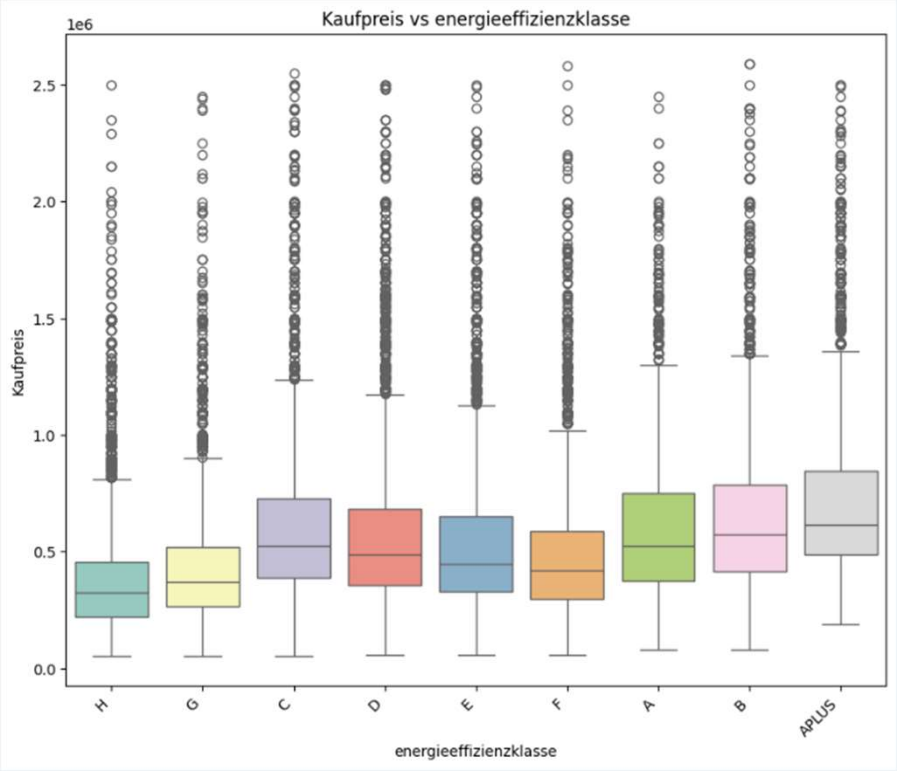
- Feature Selection
- Datenbereinigung
- Datenerkundung
- Feature Engineering
- Feature Encoding

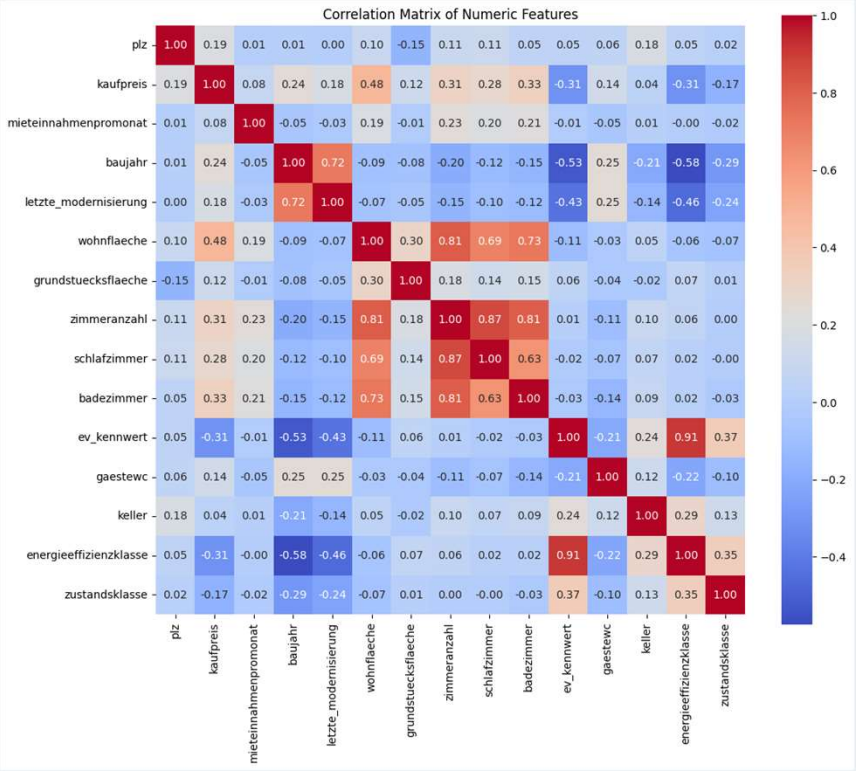
Aufbereiteter Datensatz:

- Features: 24
- Samples 43309



Datenerkundung





Ziele und Probleme des Clustering

Ziel: Erzeugung homogener Cluster mit Immobilien ähnlichen Eigenschaften und Preisspannen

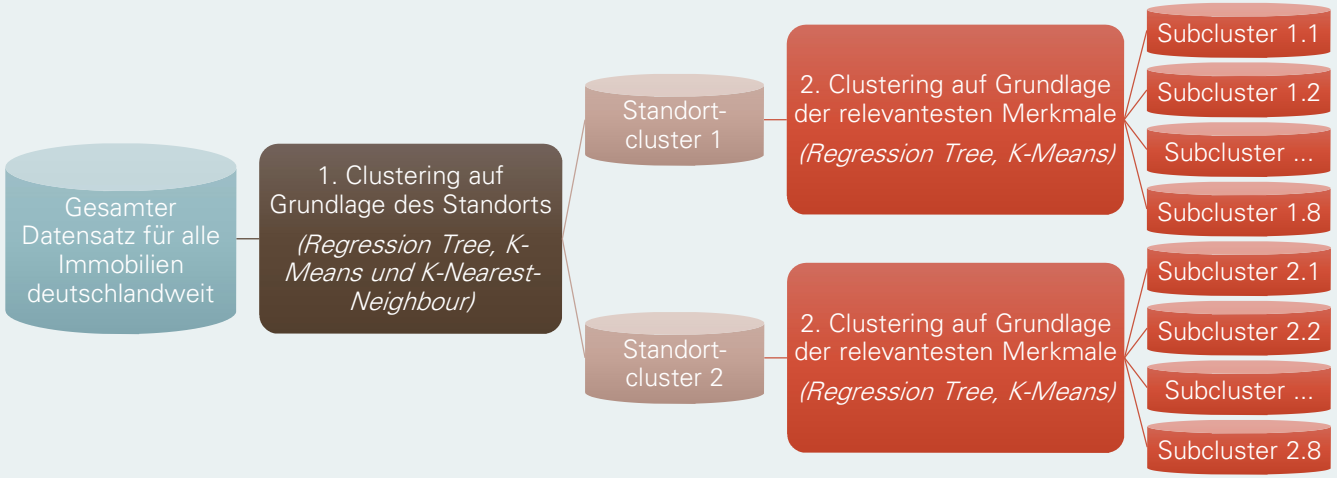
Problem: erhebliche Preisunterschiede in Abhängigkeit vom Standort

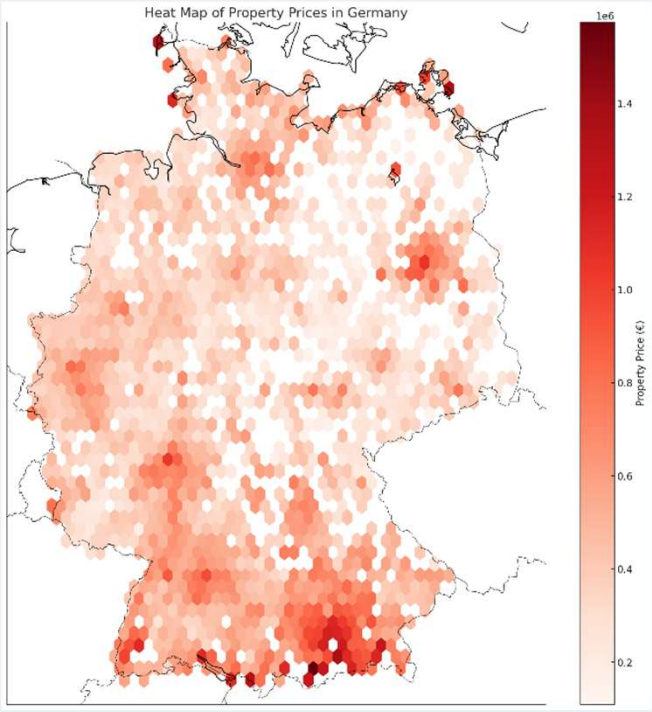
- Cluster werden hauptsächlich aufgrund des Standorts gruppiert
- Gruppierung hinsichtlich nicht standortabhängiger Merkmale
  → extreme Preisvarianz in den Clustern

Cluster-Then-Predict-Ansatz

- Reduzierung des Datensatzes auf eine Region Hamburg
- Erzeugung von 8 clustern
- Verwendete Clustering Algorithmen
    - Regression Tree (Tiefe 3)
    - K-Means (9 Features)

Cluster-Cluster-Then-Predict Ansatz





Evaluation der Clustering Ansätze

Predict Methoden

- Bewertung der Clustering Ansätze
- Identifizierung von relevanten Features
- Verwendete Algorithmen:
  - EBM
  - Lasso Regression

Bewertungsmethoden

- Tabellarische und graphische Darstellung der Cluster
- Gewichteter RMSE und MAE
- Validierung mittels Cross Validation

Ergebnisse

Cluster-Then-Predict Ansatz (Regression Tree I)

Cluster	3	4	6	7	10	11	13	14
Entscheidungsregeln								
Regel 1	Wohnfläche							
	< 225				> 225			
Regel 2	Wohnfläche				Postleitzahl			
	< 134		>134		< 21915		> 21915	
Regel 3	Wohnfläche		Postleitzahl		Wohnfläche		Längengrad	
	< 112	> 112	< 22148	> 22148	< 334	> 334	< 10,12	> 10,12

Cluster-Then-Predict Ansatz (Regression Tree II)

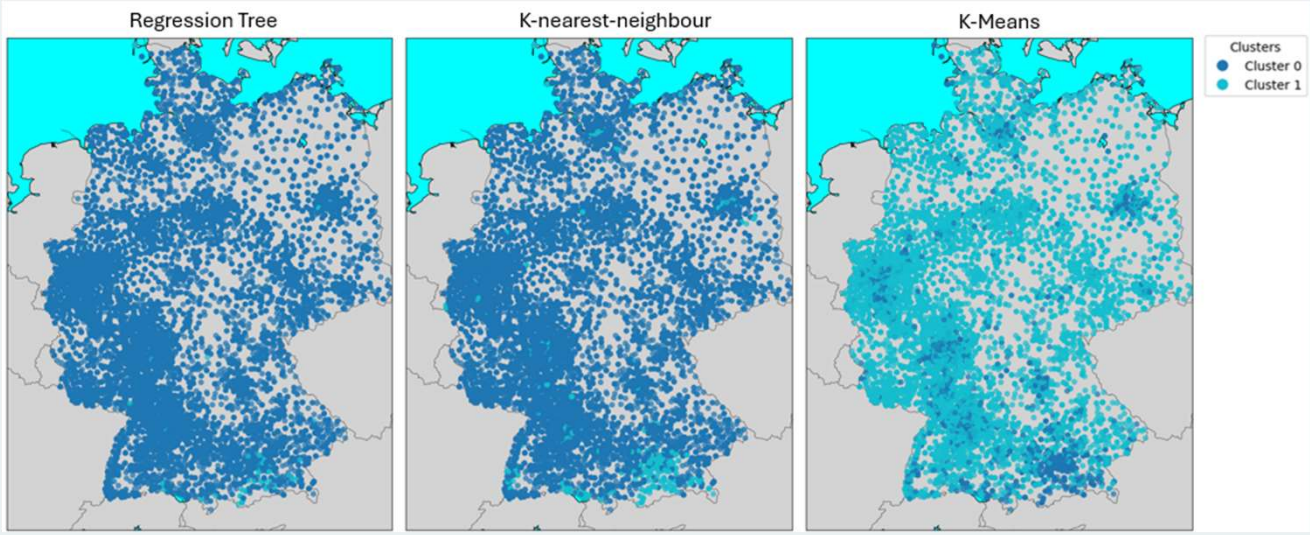
Cluster	3	4	6	7	10	11	13	14
				Mediane der Merkmale				
Postleitzahl	22117	21762	21407	22547	21357	21224	22397	22927
Baujahr	1965	1979	1978	1978	1973,5	1975,5	1970	1959
Modernisiert	1970	1984	1986	1986	1980	1979	1994	1976
Wohnfläche	98	124	163	160	260	379	289	245
Grundstücksfläche	341	489	796	649	1163	1584	1000	1295
Zimmeranzahl	4	5	6	6	8	10	9	8
Schlafzimmer	3	3	4	4	5	5	5	5
Badezimmer	1	1	2	2	3	3	3	3
EV-Kennwert	175	140	134	133	139	130	123	147
Energieeffizienzklasse	6	5	5	5	5	5	4	5
				Mediane der Zielvariable				
Kaufpreis	379000	530000	575000	772500	717000	1532500	1500000	907500

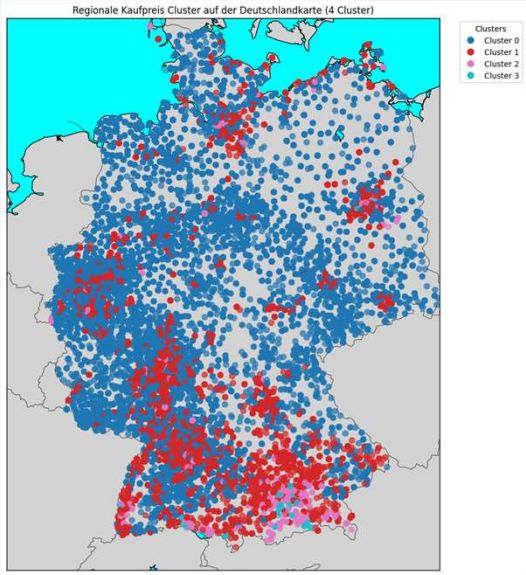
Cluster-Then-Predict Ansatz (Regression Tree III)

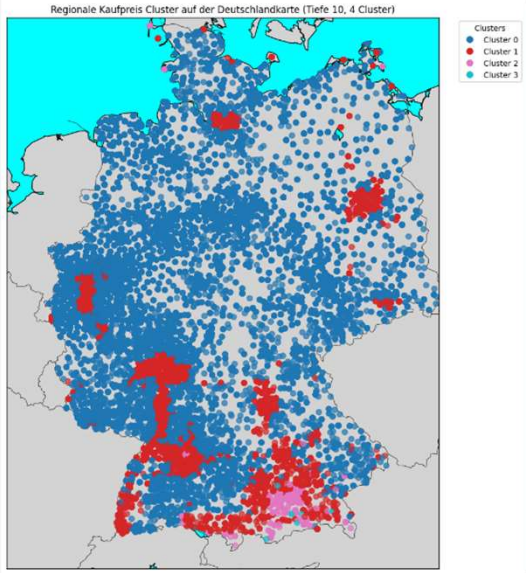
Ergebnisse Cluster-Then-Predict

Modell	Clustering Ansatz	Mittlerer MAE	Mittlerer RMSE
EBM	Ohne Clustering	120.763	179.695
	Cluster-Then-Predict (Regression Tree)	117.192	157.684
	Cluster-Then-Predict (K-Means)	138.926	179.485
Lasso Regression	Ohne Clustering	164.880	239.830
	Cluster-Then-Predict (Regression Tree)	154.078	199.139
	Cluster-Then-Predict (K-Means)	128.789	167.403

Cluster-Cluster-Then-Predict Ansatz



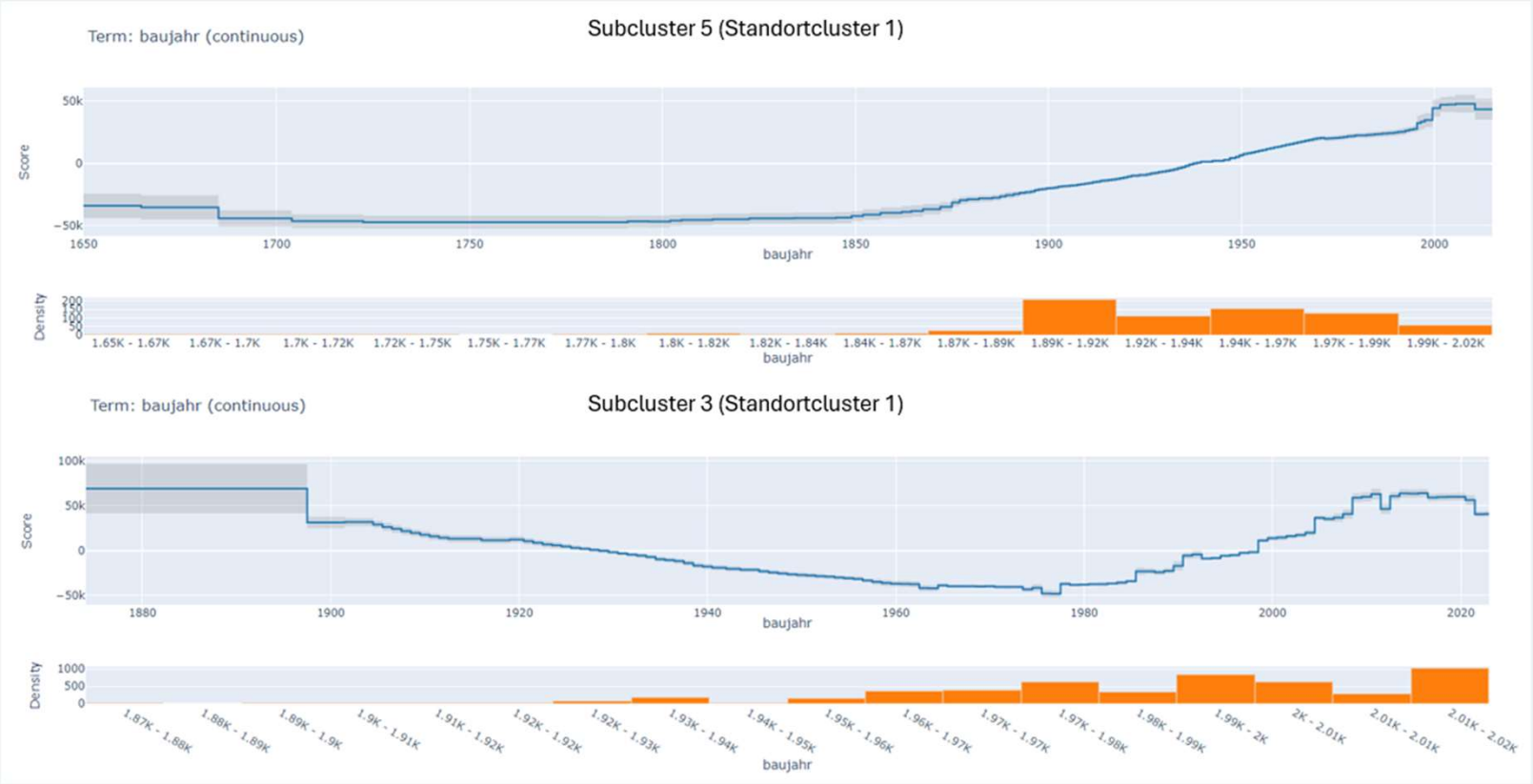




Ergebnisse Cluster-Cluster-Then-Predict

Modell	Clustering Ansatz	Mittlerer MAE	Mittlerer RMSE
EBM	Ohne Clustering	110.387	165.509
	Standortclustering (K-Means)	87.276	115.627
	Cluster-Cluster-Then-Predict (K-Means - Regression Tree)	88.696	115.291
	Cluster-Cluster-Then-Predict (K-Means - K-Means)	80.925	104.189
Lasso Regression	Ohne Clustering	170.745	244.552
	Standortclustering (K-Means)	127.272	160.076
	Cluster-Cluster-Then-Predict (K-Means - Regression Tree)	122.295	151.510
	Cluster-Cluster-Then-Predict (K-Means - K-Means)	107.996	134.297

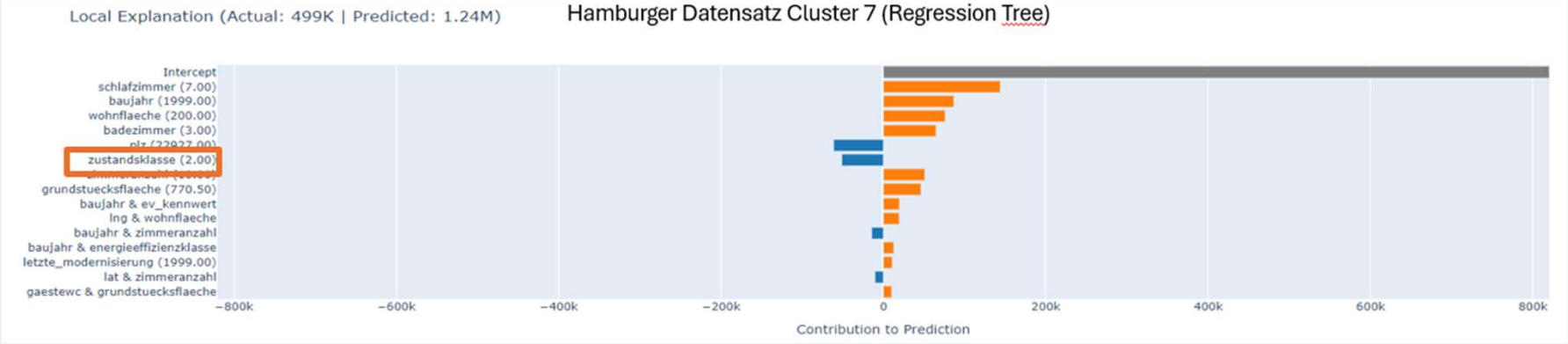
Clusterspezifische Unterschiede in den Merkmalen



Ausblick

Limitationen

- Herausforderung bei der Preisprognose von Luxusimmobilien
- Probleme bei der Berücksichtigung von Extremwerten
- Einschränkungen der Datenbasis
- Begrenzte Auswahl an Algorithmen und Modellen



Fazit

Verbesserungen

- Einfluss infrastruktureller Merkmale
- Einbeziehung von Bilddaten

Bestes Modell

- Cluster-Cluster-Then-Predict Ansatz mit dem zweifachen K-Means und dem EBM-Model

Wichtigste Feature

- Lage der Immobilie
- Bei ähnlichen Standortbedingungen: Wohnfläche, Grundstücksfläche und Baujahr

Vielen Dank für Ihre Aufmerksamkeit!