Wissenschaftlicher Schlussbericht zu Innovation Scheck (Application Number: 68795.1 INNO-SBM)

Wohnimmobilienbewertung mit Methoden des Maschinellen Lernens

Version Januar 2024

Dr. Rudolf Marty[1]

Prof. Dr. Peter Ilg [2]

[1] Rudolf Marty ist Senior wissenschaftlicher Mitarbeiter am Swiss Real Estate Institute an der HWZ Hochschule für Wirtschaft Zürich (Schweiz), Lagerstr. 5, 8021 Zürich (email-Adresse: rudolf.marty@swissrei.ch ,Telefon: ++41 43 322 26 13).

[2] Peter Ilg ist Leiter des Swiss Real Estate Instituts an der HWZ, Hochschule für Wirtschaft Zürich

Inhaltsverzeichnis

1. Einleitung

1.1 Fragestellung

In der Praxis der Wohnimmobilienbewertung in der Schweiz haben sich log-lineare hedonische Modelle als «Industriestandard» etabliert. Diese Modelle beschreiben die Logarithmen der Objektpreise als Funktion ihrer wichtigsten Objektmerkmale und Umgebungsvariablen (z.B. Entfernung zum Öffentlichen Verkehr). Mittels der Methode der kleinsten Quadrate (Ordinary Least Squares (OLS)) werden die Modellparameter des hedonischen Modelles unter Verwendung eines hinreichend grossen Datensatzes, der u.a. die Transaktionspreise enthält, geschätzt. Mit der zunehmenden Verbreitung von Verfahren des Maschinellen Lernens (ML) und dem Vorliegen von immer grösseren Datenbeständen wurden komplementär zur hedonischen Bewertungsmethode zunehmend komplexere Bewertungsverfahren eingesetzt. Ziel dieser empirischen Kurzstudie besteht darin abzuklären, ob und in welchem Masse sich die Genauigkeit von Verkaufspreisschätzungen für Eigenheime durch den Einsatz von Verfahren des Maschinellen Lernens im Vergleich zum log-linearen hedonischen Ansatz verbessern lässt.

1.2 Literatur

In der Literatur existiert bereits eine Vielzahl von Studien, die die Treffsicherheit von hedonischen Bewertungsmodellen mit ausgewählten Verfahren des Maschinellen Lernens vergleichen. Stang et al. (2022) stellten in ihrer Studie zwei hedonische Modelle einem mit einem Verfahren des Maschinellen Lernens (XGBoost-Algorithmus) berechneten Modell gegenüber, wobei bei allen Verfahren ein identischer Variablensatz zur Prognose der Objektpreise verwendet wurde. Der Datensatz umfasste 1.2 Millionen Wohnimmobilien in Deutschland. Die Studie ergab eine deutliche Überlegenheit eines Verfahrens des Maschinellen Lernens gegenüber zwei hedonischen Modellen. Auch in der Studie von Marty (2022) mit den Transaktionspreisen von Eigentumswohnungen und Einfamilienhäuser des Swiss Real Estate Datapools wurde eine Überlegenheit der mit dem Neuronalen Netzwerk-Verfahren berechneten Preise im Vergleich zu den mittels eines log-linearen hedonischen Modelles berechneten Preise festgestellt.

2. Log-lineare hedonische vs Machine Learning Modelle

2.1 Das log-lineare Regressionsmodell

Beim log-linearen hedonischen Bewertungsansatz wird der Logarithmus des Preises des i-ten Objektes, $p_i = \log(P_i)$, mit einem Vektor der für die Bewertung des i-ten Objektes relevanten k logarithmierten (quantitativen) Objektmerkmale, $\log(X_i)$,, erklärt (z.B. Alter von Objekt i):

$$(1)\ \log(P_i) = \sum_{j=1}^{k} \beta_j \ \log(X_{i,j}) + u_i$$

, wobei u_i sämtliche unsystematischen Einflüsse auf den Preis des i-ten Objektes abbildet.

Die Koeffizienten β_j geben den (isolierten) Einfluss des j-ten Objektmerkmals auf den Wert des i-ten Objektes wieder. Handelt es sich z.B. bei $X_{i,j}$ um das Alter und bei P_i um den Objektpreis des i-ten Objektes in CHF, so gibt β_j die Preissensitivität von Objekt i in Bezug auf das Gebäudealter an (um wieviel Prozent ändert sich der Transaktionspreis bei einer Veränderung des Gebäudealters um ein Prozent). Die in dieser Arbeit verwendeten Objektvariablen stammen mit einer Ausnahme (kantonale Arbeitslosenquoten von SECO) alle von der SRED-Datenbank[3]. Ein Nachteil des hedonischen Ansatzes ist, dass (nicht-lineare) Kreuz-Beziehungen zwischen den Objektmerkmalen (z.B. zwischen Objektalter und der Objektgrösse) bei einer log-linearen Modellspezifikation nicht berücksichtigt sind.

2.2 Zwei Verfahren des Maschinellen Lernens

2.2.1 Neurale Netzwerke

Ein künstliches neuronales Netzwerk ist ein zweistufiges Regressionsmodell für quantitative Erklärungsvariablen bzw. ein zweistufiges Klassifikationsmodell für kategoriale Erklärungsvariablen, das typischerweise als Netzwerkdiagramm dargestellt wird. Wird eine log-lineare Spezifikation für das hedonische Modell unterstellt, so lassen sich die Logarithmen der Objektpreise im Rahmen eines neuronalen Netzwerkes als eine Linearkombination von M Eigenschaften («features» in «hidden layers» bzw. in versteckte Schichten)
$g_m\left(\sum_{j=1}^{k} \alpha_j \ log(X_{j,i}) + \alpha_0\right)$ der Immobilie darstellen, wobei die Eigenschaften wiederum nicht-lineare Funktionen der k für die Objektbewertung relevanten Objektmerkmale sind:

$$(2)\ \log(P_i) = \sum_{m=1}^{M} g_m \left(\sum_{j=1}^{k} \omega_j \ log(X_{j,i}) + \omega_0 \right)$$

[3] Das log-lineare Modell für Einfamilienhäuser bzw. Eigentumswohnungen besteht aus 12 bzw. 13 Objekteigenschaften und einer Umgebungsvariable (kantonale Arbeitslosenquote).

Die nicht-linearen Aktivierungsfunktionen g_m werden nicht wie die Parameter α_j geschätzt, sondern a priori spezifiziert, z.B. durch die Sigmoid-Funktion $g=1/(1+e^{v_j})$, $v_j=\alpha_0 + \alpha_j\, log(X_{j,i})$[4]. Die Koeffizienten ω_j, j=1,…,k bzw. ω_0 (die auch Gewichts- bzw. Verzerrungskoeffizienten genannt werden) müssen durch die Minimierung der Summe der Abstandsquadrate der theoretischen von den beobachteten Werten mittels iterativer Verfahren (stochastic Gradient Descent-Methode) geschätzt werden.

2.2.2 Der Extreme Gradient Boosting - Algorithmus

Der «Extreme Gradient Boosting»-Algorithmus h ist ein auf einem Entscheidungsbaum-Verfahren basierender Algorithmus, der eine Vielzahl von sogenannt «schwachen» Entscheidungsbäumen h_m kombiniert:

(3) $h(P_i \mid x_{i,1}, \ldots, x_{i,k}) = \sum_{j=1}^{M} u_m\, h_m\, (P_i \mid x_{i,j})$

, wobei $x_{i,j}$ das j-te Objektmerkmal des i-ten Objektes und $h_m(.)$ der m-te (schwache) Entscheidungsbaum von insgesamt M schwachen Entscheidungsbäumen ist. Für eine intuitive Erklärung des XGBoost-Algorithmus siehe auch Stang et al. (2022).

Der XGBoost-Algorithmus benützt das «Gradient Descent»-Verfahren durch das Hinzufügen zusätzlicher schwacher Entscheidungsbäume für die Minimierung der Funktion in (3).

3. Der verwendete Datensatz und Prognosevergleich

3.1 SRED-Datensatz (Transaktionspreise und Objektmerkmale)

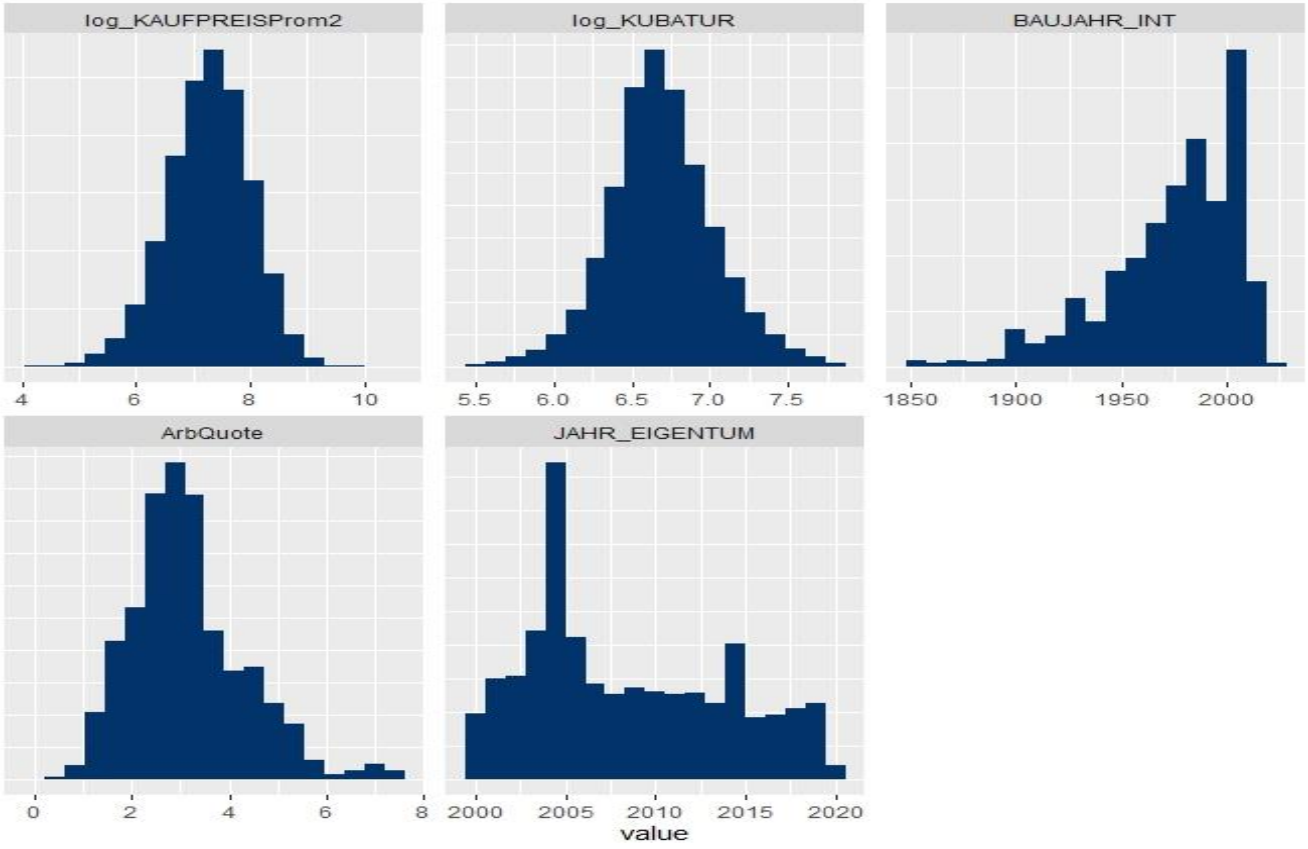
Die Transaktionspreise der Wohneigentumsobjekte und ihre Objektmerkmale stammen alle aus der SRED-Datenbank[5] der Periode 1. Quartal 2000 – 1. Quartal 2020. Bei den Preisen handelt es sich ausschliesslich um Transaktionspreise aus Freihandverkäufen von Objekten (EFH: Einfamilienhäuser; EGTW: Eigentumswohnungen) in der gesamten Schweiz. Die Objekte sind nicht geocodiert, d.h. ihr Standort kann nur bis auf ihre Postleitzahl identifiziert werden. In den Grafiken 1 und 2 sind die wichtigsten univariaten Statistiken der Verteilungen der Objektpreise und der stetigen und diskreten bzw. ordinalen Objektmerkmale aufgeführt.

[4] Zur Schätzung des Neuronalen Netzwerkes wurde im Statistikpaket R die Funktion «net» verwendet.

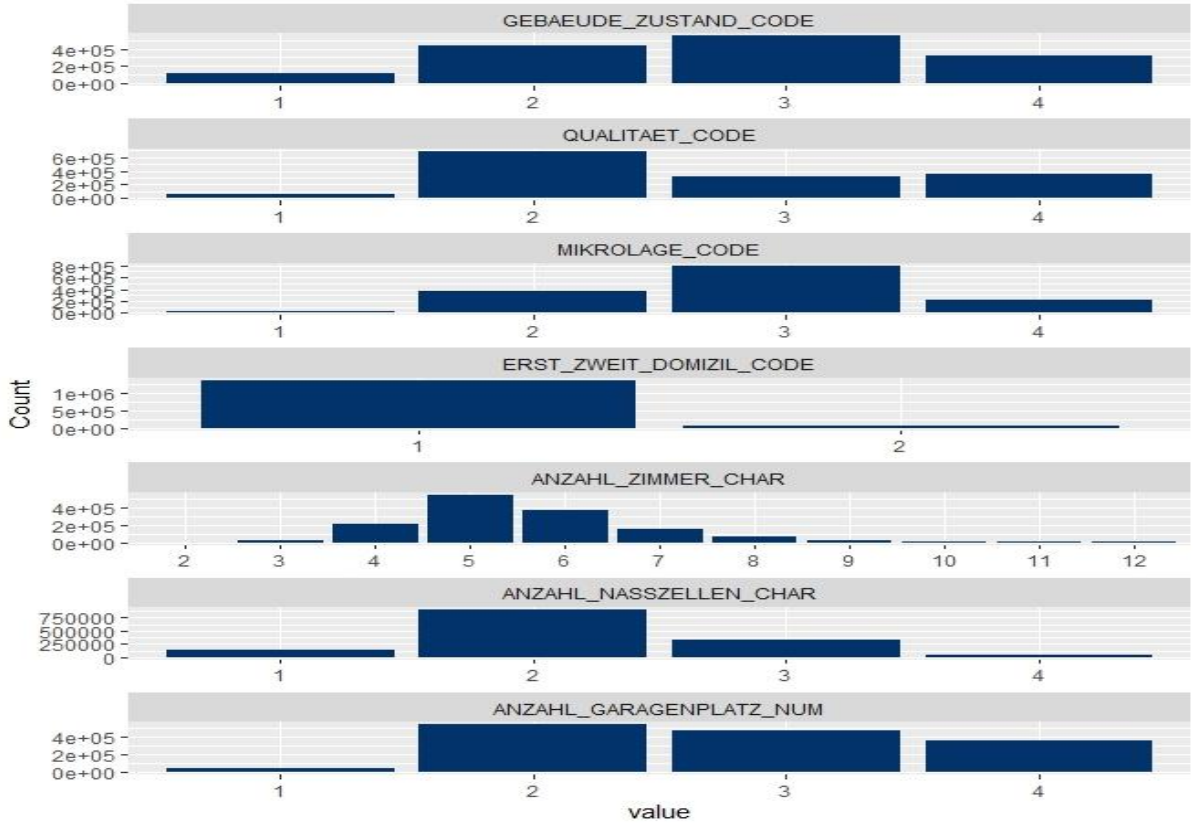
[5] Der Swiss Real Estate Datapool ist ein Verein, dessen Ziel die Förderung von Markteffizienz und -transparenz im Schweizer Eigenheimmarkt durch das Pooling von Immobilientransaktionsdaten ist.

Die Arbeitslosenquote ist nicht im Datenpool. Sie wurde von einer externen Quelle (Seco) beigefügt.

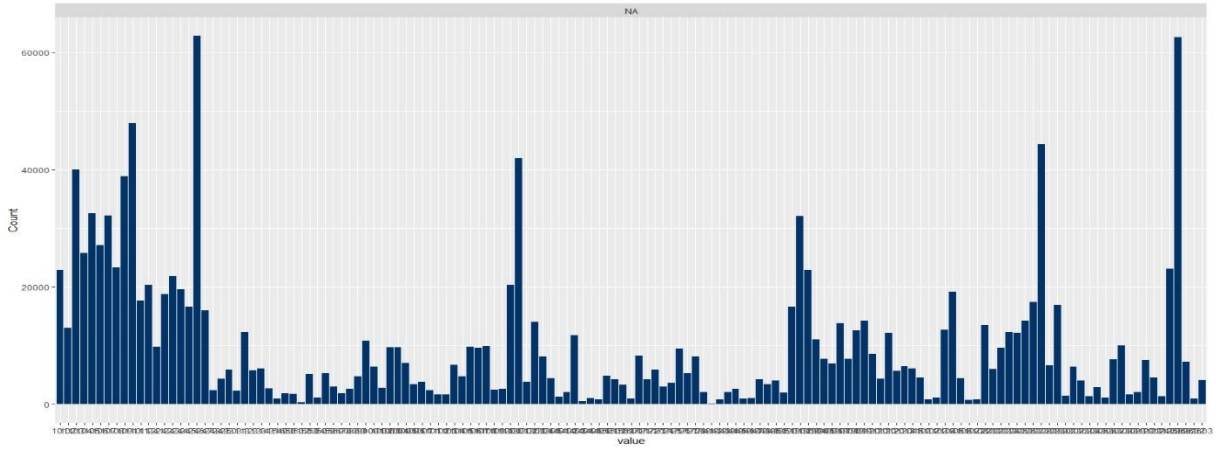
Grafik 1a: EFH-Transaktionspreise und ihre stetigen Merkmale; Periode: 2000 Q1-2020 Q1 (Quelle: SRED, Seco)



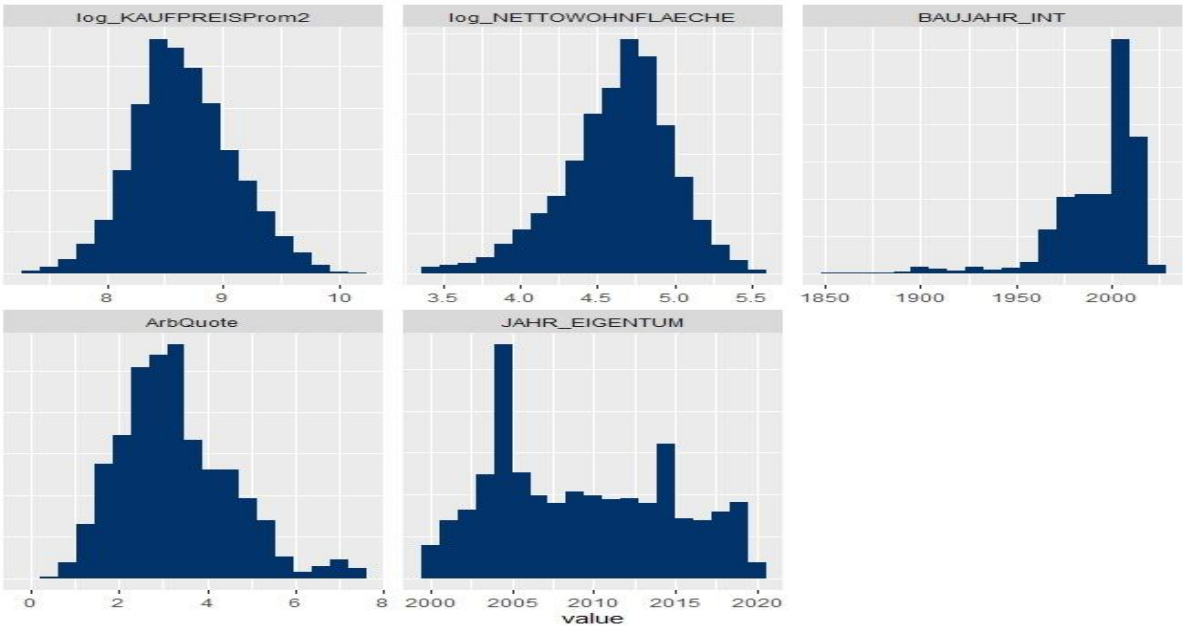
Grafik 1b: Ordinale und diskrete Objektmerkmale, Periode: 2000 Q1 – 2020 Q1 (Quelle: SRED)



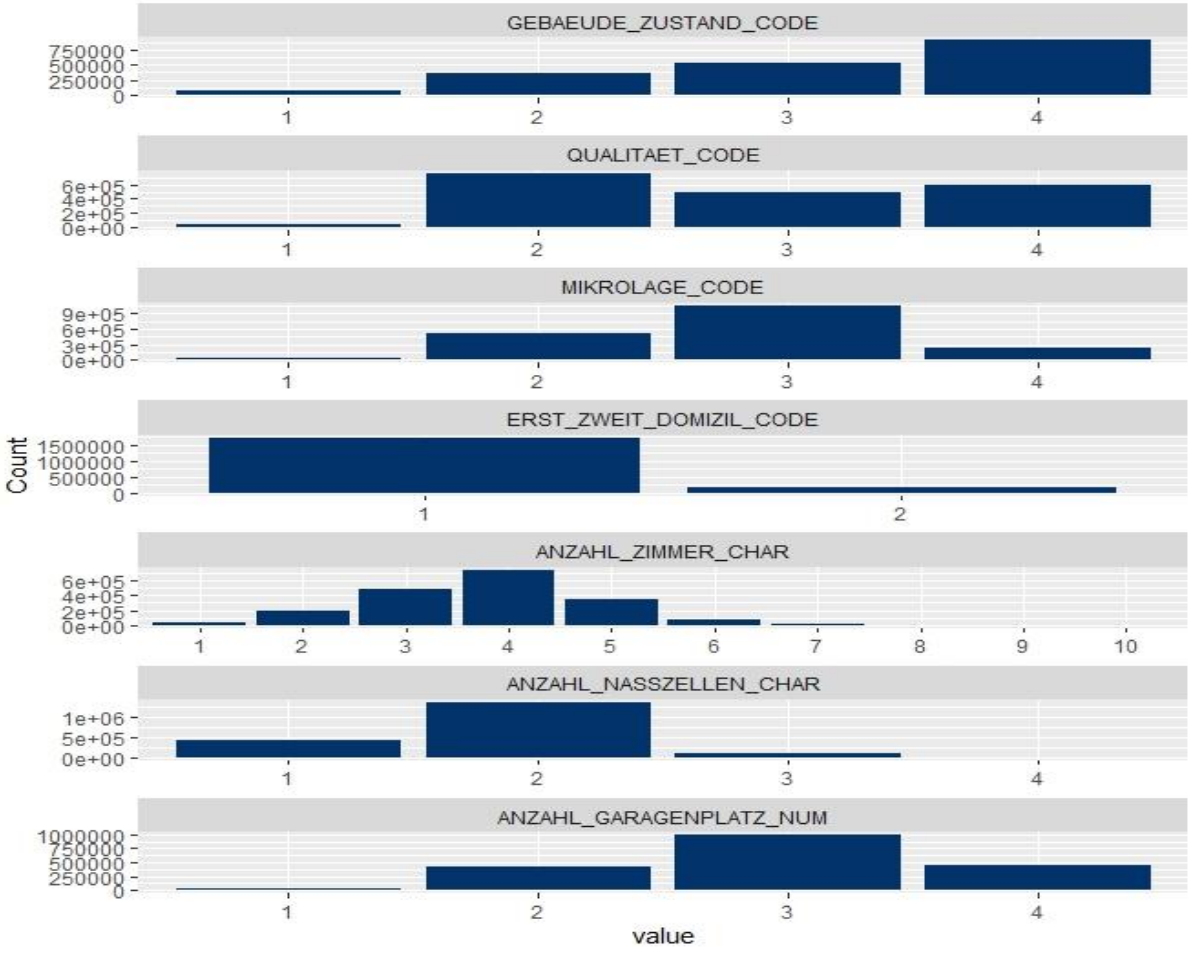
Grafik 1c: Standortbezirke Einfamilienhäuser (total: 144, als Faktor-Variable), Periode: 2000 Q1 – 2020 Q1 (Quelle: SRED)



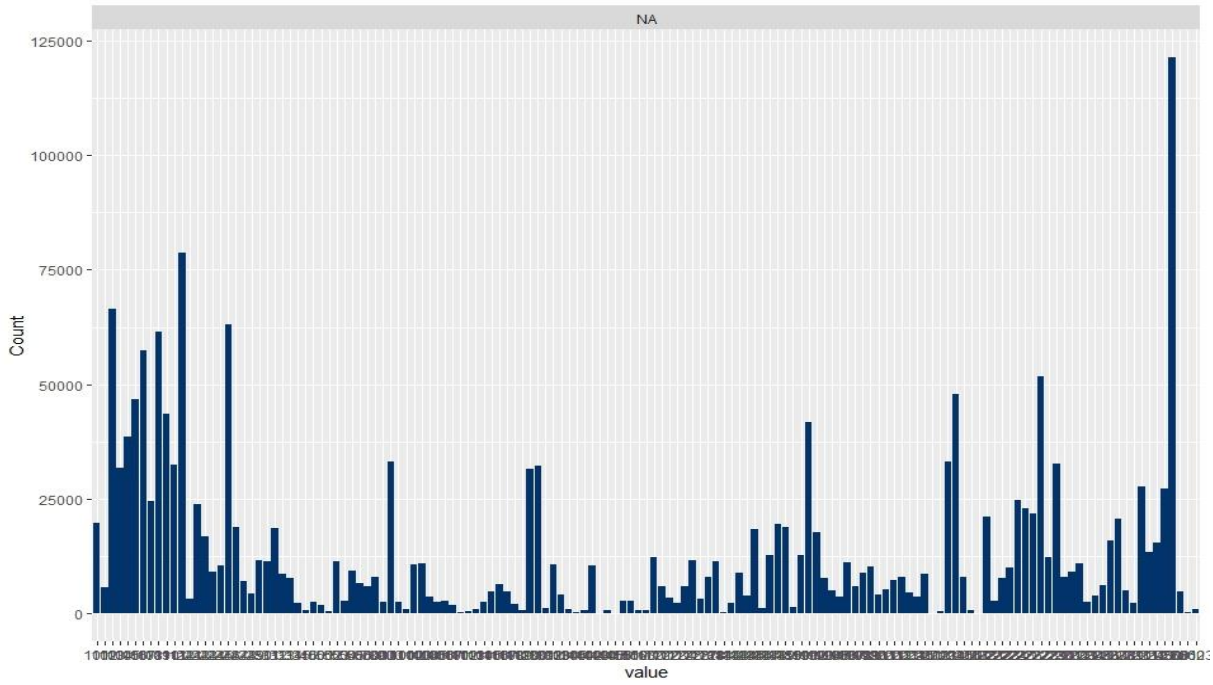
Grafik 2a: EGTW-Transaktionspreise und ihre stetigen Merkmale; Periode: 2000 Q1 – 2020 Q1 (Quelle SRED, Seco)



Grafik 2b: Ordinale und diskrete Objektmerkmale, Periode: 2000 Q1 – 2020 Q1 (Quelle: SRED)



Grafik 2c: Standortbezirke Eigentumswohnungen (total: 144, als Faktor_Variable), Periode: 2000 Q1 – 2020 Q1 (Quelle: SRED)



Gemäss den Grafiken 1 bzw. 2 bezieht sich die Median-Transaktion auf ein 1981 bzw. 2001 erstelltes Objekt mit 5 bzw. 4 Zimmern, knapp 800 m^3 bzw. 107 m^2 Kubatur bzw. Wohnfläche, das zu einem Preis von CHF 760'000 bzw. CHF 600'000 den Besitzer wechselte. Bei den Eigentumswohnungen lag der Preis pro m^2 Nettowohnfläche zwischen CHF 1'500 (Minimum) und CHF 25'000 (Maximum), wobei der Medianwert CHF 5'600 betrug.

3.2 Prognosevergleich hedonisches Modell vs zwei ML-Verfahren

Als Umgebungsvariablen der gehandelten Wohneigentumsobjekte steht die kantonale Arbeitslosenquote (auf jährlicher Basis; Quelle: Seco) zur Verfügung.

Tabelle 4:Prognosefehler log-lineares hedonisches Modell für Eigentumswohnungen u. Einfamilienhäuser

$\log(P_i) = \sum_{j=1}^{k} \beta_j \log(X_{i,j}) + u_i$ Eigentumswohnung (EGTW)			$\log(P_i) = \sum_{j=1}^{k} \beta_j \log(X_{i,j}) + u_i$ Einfamilienhaus (EGTW)		
Teststatistiken	**Preis standardisiert** (pro m^2 Wohnfläche)	**Preis nicht standardisiert**	**Teststatistiken**	**Preis standardisiert** (pro m^2 Grundstückfläche)	**Preis nicht standardisiert**
Trainingsdaten (N=125'546)			**Trainingsdaten (N=94'767)**		
RMSE	0.2311	0.2338	RMSE	0.4852	0.2450
MAE	0.1759	0.1763	MAE	0.3636	0.1750

R^2	0.7162	0.7100	R^2	0.5495	0.8791
Testdaten (N=31'386)			**Testdaten (N=23'669)**		
RMSE	**0.2331**	**0.2319**	**RMSE**	**0.4913**	**0.2514**
MAE	**0.1780**	**0.1769**	**MAE**	**0.3731**	**0.1836**
Innerhalb10%	**0.3691**	**0.3722**	**Innerhalb10%**	**0.1849**	**0.3732**
Innerhalb20%	**0.6553**	**0.6576**	**Innerhalb20%**	**0.3609**	**0.6545**
Erklärung: MAE: Mittelwert der absoluten Preisfehler RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler					

Quelle: eigene Berechnung, SRED-Datenbank, SECO

Entsprechend den in der Tabelle 5 dokumentierten Prognosefehler-Kennzahlen kann mittels des ANN-Verfahrens die Treffsicherheit der theoretischen Objektpreise im Vergleich zu den hedonischen Modellpreisen um 3-4 Prozentpunkte erhöht werden (in Bezug auf die Kenngrössen Innerhalb10% bzw. Innerhalb20%).

Tabelle 5a, 5b: P_i tatsächlich. vs. P_i berechnet, EGTW, EFH, Neuronales Netzwerk (6 «hidden layers»)

$\log(P_i) = \sum_{m=1}^{M} g_m \left(\sum_{j=1}^{k} \omega_j \, log(X_{j,i}) + \omega_0 \right)$ Eigentumswohnung (EGTW)			$\log(P_i) = \sum_{m=1}^{M} g_m \left(\sum_{j=1}^{k} \omega_j \, log(X_{j,i}) + \omega_0 \right)$ Einfamilienhaus (EFH)		
Teststatistiken	**Preis standardisiert** (pro m^2 Wohnfläche)	**Preis nicht standardisiert**	**Teststatistiken**	**Preis standardisiert** (pro m^2 Grundstückfläche)	**Preis nicht standardisiert**
Trainingsdaten (N=125'546)			**Trainingsdaten (N=94'767)**		
RMSE	0.0780	0.0770	RMSE	*0.0715*	0.0370
MAE	0.0588	0.0577	MAE	0.0540	0.0266
R^2	0.9677	0.9694	R^2	*0.5389*	*0.7571*
Testdaten (N=31'386)			**Testdaten (N=23'669)**		
RMSE	**0.2431**	**0.2380**	**RMSE**	**0.4587**	**0.2379**
MAE	**0.1657**	**0.1615**	**MAE**	**0.3476**	**0.1709**
Innerhalb10%	**0.3992**	**0.4141**	**Innerhalb10%**	**0.1979**	**0.4040**
Innerhalb20%	**0.6899**	**0.7008**	**Innerhalb20%**	**0.3883**	**0.6831**
Erklärung: MAE: Mittelwert der absoluten Preisfehler RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler					

Quelle: eigene Berechnung, SRED-Datenbank, SECO

Tabelle 6a, 6b: P_i tatsächlich. vs. P_i berechnet, EGTW, EFH, XGBoost-Verfahren

$h(P_i \mid x_{i,1}, \ldots, x_{i,k}) = \sum_{j=1}^{M} u_m h_m (P_i \mid x_{i,j})$ Eigentumswohnung (EGTW)			$h(P_i \mid x_{i,1}, \ldots, x_{i,k}) = \sum_{j=1}^{M} u_m h_m (P_i \mid x_{i,j})$ Einfamilienhaus (EFH)		
Zu prognostizierende Variable	**Preis standardisiert (pro m^2)**	**Preis nicht standardisiert**	**Zu prognostizierende Variable**	**Preis standardisiert (pro m^2)**	**Preis nicht standardisiert**
Trainingsdaten (N=125'546)			**Trainingsdaten (N=94'767)**		
RMSE	0.1879	0.1771	RMSE	0.3713	0.1764
MAE	0.1421	0.1334	MAE	0.2812	0.1311
R^2	0.8106	0.8320	R^2	0.6121	0.9399
Testdaten (N=31'386			**Testdaten (N=23'669)**		
RMSE	**0.2189**	**0.2031**	**RMSE**	**0.4462**	**0.2294**
MAE	**0.1586**	**0.1520**	**MAE**	**0.3337**	**0.1638**
Innerhalb10%	**0.4228**	**0.4376**	**Innerhalb10%**	**0.2144**	**0.4279**
Innerhalb20%	**0.7129**	**0.7277**	**Innerhalb20%**	**0.4026**	**0.7110**
Erklärung: MAE: Mittelwert der absoluten Preisfehler RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler					

Quelle: eigene Berechnung, SRED-Datenbank, SECO

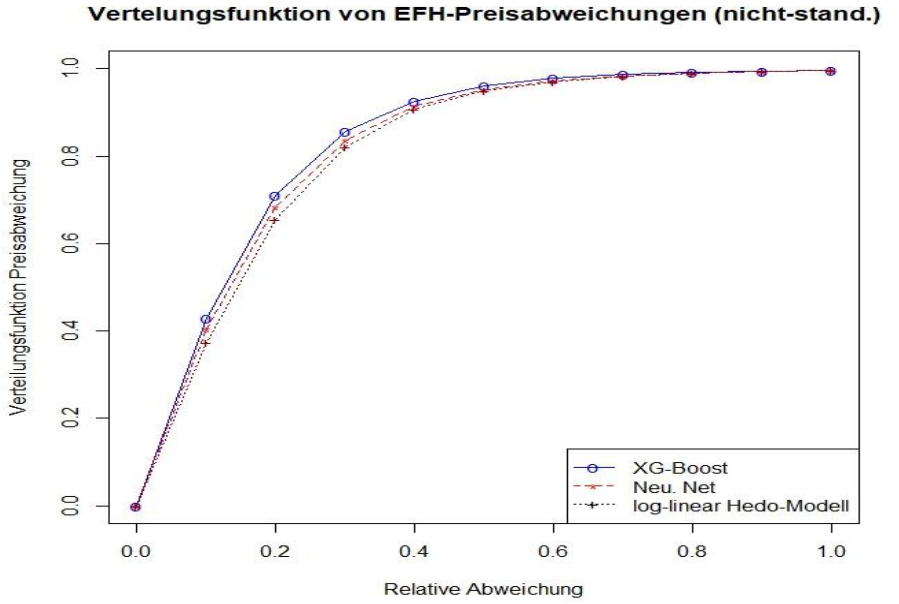
Zusammenfassend kann festgestellt werden, dass die Prognosegenauigkeit der mittels eines Neuronalen Netzwerkes berechneten (theoretischen) Objektpreise entsprechend den Kriterien «Innerhalb10%» bzw. «Innerhalb20%» um 4%-Punkte im Vergleich zu einem log-linearen hedonischen Modell erhöht wird. Durch die Anwendung des XGB-Algorithmus kann die Treffsicherheit der theoretischen XGB-Preise zusätzlich gegenüber den Neuronalen Netzwerken um 3%-Punkte erhöht werden.

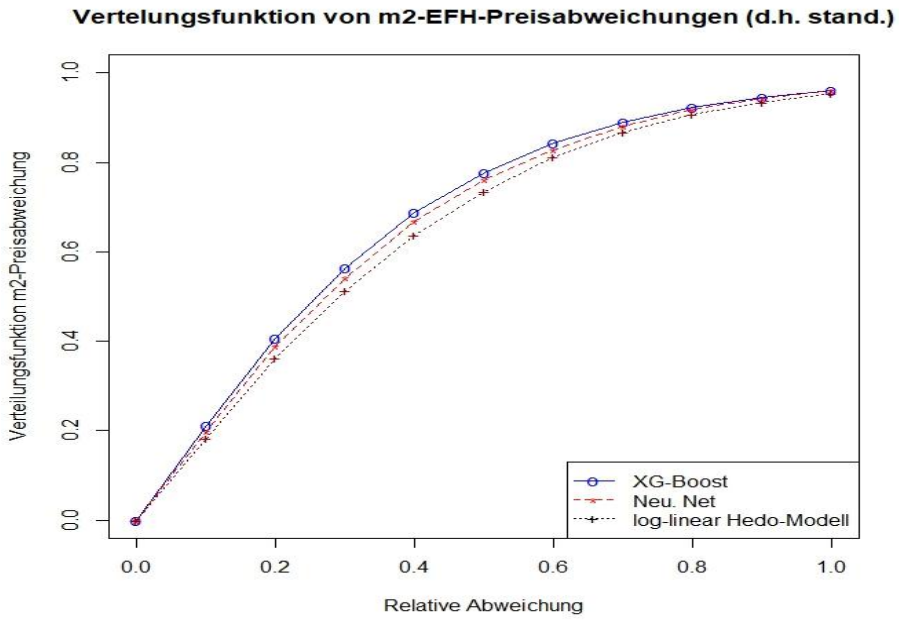
4. Fazit: Prognosegüte hedonisches Modell vs zwei ML-Verfahren

Sowohl bei EGTWs als auch bei EFHs erwies sich erstens der XGBoost-Algorithmus als überlegen im Vergleich zum Neuralem Netzwerk-Algorithmus und vor allem im Vergleich zu den log-linearen Modellen. Innerhalb eines Intervalls von +/- 10% um den Transaktionspreis lagen 44% (bei EGTWs) bzw. 43% (bei EFHs) aller theoretischen EGTW- bzw. EFH-Preise, die mittels des XGBoost-Verfahrens geschätzt wurden. Bei den log-linearen Modellen betrugen die entsprechenden Werte nur 37%. Zweitens spielt es bei EGTWs fast keine Rolle, ob theoretische m^2-Preise oder theoretische Objektpreise mittels ihrer Objektmerkmale berechnet werden, sind doch die Prognosefehler der standardisierten theoretischen Preise nur unwesentlich höher im Vergleich zu den theoretischen Objektpreisen. Bei EFHs lassen sich die

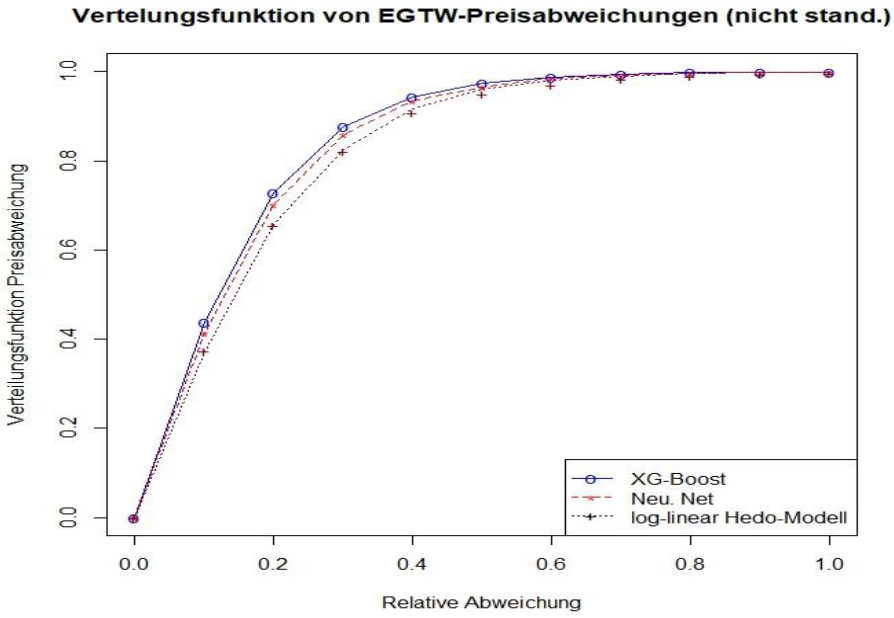
Objektpreise gemäss den Ergebnissen in der Tabelle 5b bzw. 6b deutlich präziser schätzen als die entsprechenden m^2-Preise.

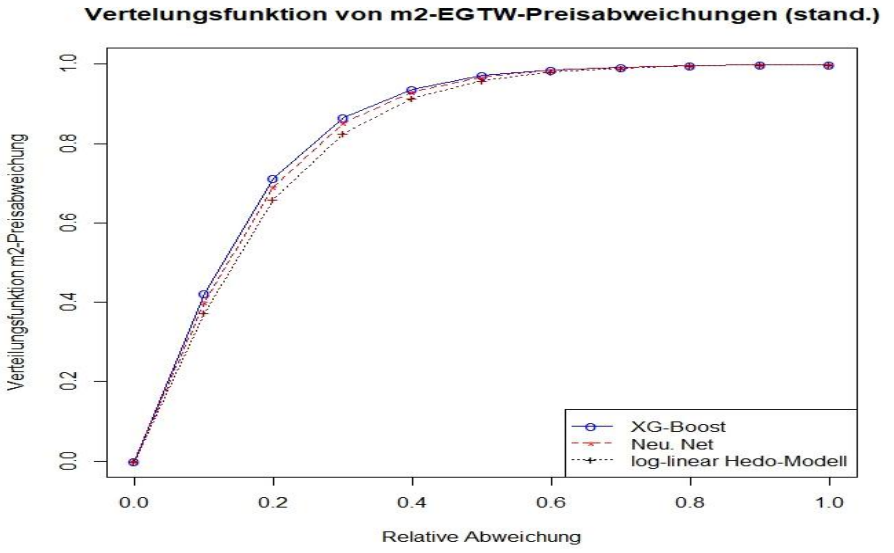
Grafik 3a und 3b: Verteilungsfunktion der relativen Abweichung des EFH-Transaktionspreise von theoretischen reisen (nicht standardisiert und standardisiert), Testdaten (N=22'669)



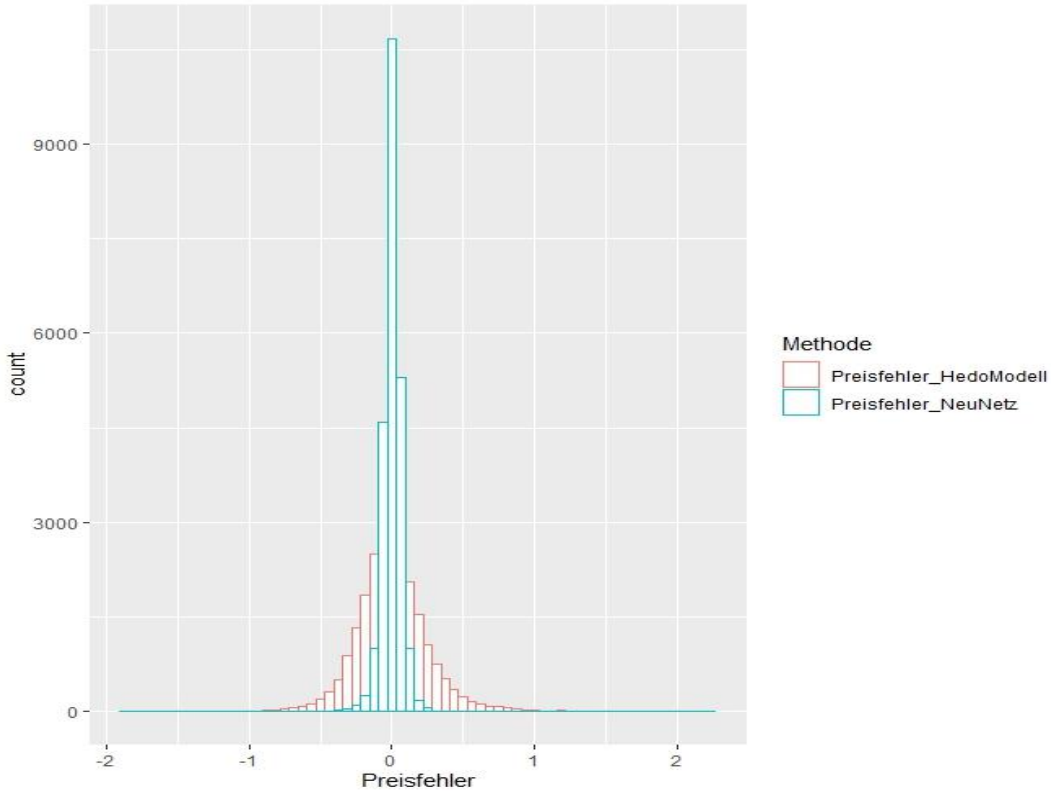


Grafik 3c und 3d: Verteilungsfunktion der relativen Abweichung des EGTW-Transaktionspreise von theoretischen Preisen (nicht standardisiert und standardisiert), Testdaten (N=31'386)





Grafik 4a: Histogramm der relativen Abweichung des EFH-Transaktionspreise von theoretischen Preisen (nicht standardisiert), Testdaten (N=22'669)



Anhang: Quantifizierung der Prognosefehler

Wird mit $P_i^{tatsächlich}$ (bzw. $\log(P_i^{tatsächlich})$) bzw. mit $P_t^{berechnet}$ (bzw. $\log(P_i^{tatsächlich})$) der Transaktionspreis bzw. dessen Logarithmus bzw. der berechnete Preis bzw. dessen Logarithmus des i-ten Objektes bezeichnet, so ist der Mittelwert des absoluten Preisfehlers (MAE) innerhalb einer Stichprobe mit Beobachtungsumfang N folgendermassen definiert:

$$MAE = \sum_{k=1}^{N} (\left| P_i^{berechnet} - P_t^{tatsächlich} \right|)$$

Der Wurzel aus dem Mittelwert des quadratischen Preisfehlers ist durch folgenden Ausdruck gegeben:

$$RMSE = (\sum_{k=1}^{N} (P_i^{berechnet} - P_t^{tatsächlich})^2)^{0.5}$$

Für die logarithmierten Preise geltend die analogen Definitionen, wobei MAE bzw. RMSE als mittlere absolute bzw. Wurzel aus der mittleren quadratischen (relativen) Preisabweichungen interpretiert werden können[6].

InnerhalbXX% gibt den Anteil der berechneten Preise an der Stichprobe an, der eine maximale (absolute) Abweichung von XX% aufweist:

$$Innerhalb10\% = f(\left| P_i^{berechnet} - P_t^{tatsächlich} \right| < 0.1 * P_t^{tatsächlich})/N,$$

wobei f(.) die absolute Häufigkeit wiedergibt.

[6] Der Ausdruck $\log(P_i^{berehcnet}) - \log(P_i^{tatsächlich})$ kann als $1+x^{rel.Diffferenz}$ ausgedrückt werden ($x^{rel.Diffferenz}$: relative Abweichung des berechneten vom tatsächlichen Preis) für hinreichend kleine x wegen der Taylor-Regel, gemäss der gilt: $\log(1+x) \approx x$.

5. Literatur

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Predictions.* Standford: Springer Series in Statistics.

Marty, R. (November 2022). Mit künstlichen Neuronalen Netzwerken Eigenheimpreise genauer schätzen als mit klassischen Hedomodellen? *Swiss Real Estate Journal No. 25*.

Stang, M., Krämer, B., Nagl, C., & Schäfers, W. (2022). From human business to machine learning- methods for automating real estate appraisals and their practical implications. *Zeitschrift für Immobilienökonomie 9(4)*, 1-28.