HWZ Working Paper Series

Machine-Learning-Verfahren versus hedonische Bewertung von Wohnimmobilien

2021

Eine explorative Analyse

Dr. Rudolf Marty

Hochschule für Wirtschaft Zürich

HWZ

HWZ
Die Hochschule für Wirtschaft
in Zürich

Machine-Learning-Verfahren versus hedonische Bewertung von Wohnimmobilien

Eine explorative Analyse[1]

Rudolf Marty[2]

[1] Wertvolle Kommentare zu einer früheren Version dieser Arbeit stammen von Cedric Bührer, die an dieser Stelle verdankt werden.

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

HWZ
Die Hochschule für Wirtschaft
in Zürich

Abstract

Ziel dieser Studie ist der Vergleich der Prognosefehler hedonischer Modelle mit denjenigen von drei populären Verfahren des maschinellen Lernens (ML), die zunehmend zur Bewertung von Wohneigentum (Einfamilienhäusern, Eigentumswohnungen) eingesetzt werden. Bei den verwendeten Daten handelt es sich um die Transaktionspreise von ca. 120'000 Einfamilienhäusern und 160'000 Eigentumswohnungen sowie deren wichtigste Objektmerkmale, die im Zeitraum 2000-2020 in der gesamten Schweiz erhoben wurden (Quelle: SRED-Datensatz). In Übereinstimmung mit den Resultaten anderer Studien mit Schweizer und US-Immobiliendaten zeigt sich auch in dieser Studie bei zwei von drei berücksichtigten ML-Verfahren eine klare Überlegenheit dieser Methoden im Vergleich zu hedonischen Bewertungsverfahren (bei einer log-linearen Spezifikation der Bewertungsgleichung).

Schlüsselwörter Hedonische Immobilienbewertung, Machine Learning, Robuste Regression, Random Forest, Gradient Boosting, Artificial Neuronal Network

Reviewed Dezember 2020

Verfügbar online Februar 2021

HWZ
Die Hochschule für Wirtschaft
in Zürich

Inhaltsverzeichnis

1. Einleitung

1.1 Fragestellung

In der Immobilienbewertung hat die hedonische Methode vor allem bei der Bewertung von Wohnimmobilien in der Schweiz eine rasch wachsende Bedeutung erfahren. Das Standardverfahren besteht in der Spezifikation eines linearen bzw. log-linearen Modells, das die Objektpreise bzw. deren Logarithmen als Funktion der wichtigsten Objektmerkmale und ihrer Umgebungsvariablen (z. B. Steuersatz Standortgemeinde, Aussenlärm etc.) beschreibt. Mittels der Methode der kleinsten Quadrate (Ordinary Least Squares (OLS)) werden die Modellparameter des hedonischen Modells unter Verwendung eines hinreichend grossen Datensatzes, der u. a. die Transaktionspreise der Objekte enthält, geschätzt. Mit Hilfe dieser Parameter können schliesslich Objektbewertungen vorgenommen werden, d. h. es kann für ein beliebiges Objekt ein theoretischer Preis berechnet bzw. prognostiziert werden, der im Mittel dem tatsächlichen bzw. dem Transaktionspreis entspricht, sollte für dieses Objekt effektiv eine Transaktion stattfinden. Für eine Übersicht über die hedonische Bewertung von Immobilien siehe z. B. Sirmans et al. (2005). Mit der zunehmenden Verbreitung von Verfahren des Maschinellen Lernens (ML) und dem Vorliegen von immer grösseren Datenbeständen wurden komplementär zur hedonischen Bewertungsmethode auch Verfahren des ML zur Prognose von Immobilienpreisen eingesetzt, darunter die Methode der künstlichen neuronalen Netze (ANN, Hastie et al. (2009)) und des Random Forests (RF, siehe Breimann (2001)) sowie des Gradient Boosting (GB, siehe Friedman (1999)). Ziel dieser explorativen empirischen Untersuchung besteht darin abzuklären, ob und in welchem Masse sich die Prognosequalität von Wohneigentumspreisen durch den Einsatz dieser drei ausgewählten Verfahren des ML im Vergleich zum hedonischen Ansatz verbessern lässt.

1.2 Literatur

Vor allem in der angelsächsischen Literatur existiert bereits eine Vielzahl von Studien, die die Preis- bzw. Bewertungsfehler von hedonischen Bewertungsmodellen mit denjenigen ausgewählter ML-Verfahren (RF, GB, XG Boost) vergleichen. Kok und Martinez-Barbose (2017) stellten in ihrer Studie ein hedonisches Modell drei ML-Verfahren gegenüber, wobei bei allen Verfahren ein identischer Variablensatz zur Prognose der Objektpreise verwendet wurde (Objektmerkmale, Umgebungs- und makroökonomische Grössen). Der Datensatz umfasste gut 5'000 Renditeimmobilien (d. h. überwiegend Mehrfamilienhäuser) der Periode 2011-2016, die mehrheitlich aus den drei Bundesstaaten Kalifornien, Florida und Texas stammten. Die Studie

ergab für drei Modellspezifikationen eine generelle Überlegenheit der ML-Verfahren, wobei sich die GB-Methode bei zwei von drei Modellspezifikationen als die Methode mit dem minimalen Preisfehler erwies.

Mittels Daten des Schweizer Immobilienmarktes verglichen Sconamiglio et al. (2019) die Prognosequalität eines konventionellen hedonischen Modells für die Preise von Einfamilienhäusern mit den Preisfehlern von drei ausgewählten Verfahren des ML (ANN, RF, GB). Ihr Datensatz umfasste die Preise und wichtigsten Objektmerkmale sowie die Umgebungsvariablen von 123'000 Einfamilienhäusern in der gesamten Schweiz im Zeitraum 2005-2017. Der zentrale Befund von Sconamiglio et al. (2019) lautet, dass das GB-Verfahren unter den insgesamt sechs eingesetzten Schätzmethoden (Kleinstquadrat- bzw. OLS-Methode, robuste Schätzung, Mixed-Effect-Verfahren, ANN, GB, RF) entsprechend fünf von sechs Teststatistiken (Wurzel aus mittlerem quadratischem Fehler (RMSE), mittlerem absolutem Fehler (MAE), Median des absoluten Fehlers, Innerhalb10%, Innerhalb20%[3]) mit Abstand die tiefsten Prognosefehler generierte. Zwei von drei Schätzverfahren (OLS, Robustes Verfahren), denen der hedonische Ansatz zugrunde liegt, wiesen gemäss vier von sechs Teststatistiken die grössten Prognosefehler auf[4].

2. Theorie: Hedonischer Ansatz vs. Machine-Learning-Verfahren

2.1 OLS-Schätzung von hedonischem Ansatz

Beim hedonischen Bewertungsansatz wird der Preis P_i des i-ten Objektes bzw. 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 (quantitativen) Objektmerkmale X_i bzw. mit den Logarithmen der k relevanten (quantitativen) Objektmerkmale, $\log(X_i)$, erklärt (z. B. Alter von Objekt i):

$$(1a) \quad P_i = \sum_{j=1}^{k} \beta_{j,i} \, X_{j,i} + u_i \qquad (1b) \quad \log(P_i) = \sum_{j=1}^{k} \beta_{j,i} \, \log(X_{j,i}) + u_i$$

u_i bildet hierbei sämtliche unsystematischen Einflüsse auf den Preis des i-ten Objektes ab.

[3] InnerhalbXX% gibt den Anteil der berechneten Preise an der Stichprobe an, der eine maximale absolute Abweichung vom Transaktionspreis von XX % aufweist.

[4] Die Studie von Sconamiglio et al. (2019) spezifiziert u. a. ein log-lineares Modell für die Einfamilienhauspreise, d. h. die Preisfehler des Modells können approximativ als prozentuale Fehler interpretiert werden. Für eine rollende Schätzperiode des Modells erhielten Sconamiligo et al. (2019) für die OLS- bzw. robuste Schätzmethode einen mittleren absoluten Prognosefehler von 18.7 % bzw. 18.6 % (Die Wurzel aus dem mittleren quadrierten Prognosefehler betrug 26.4 % bzw. 26.5 %).

Der Vorteil des hedonischen Ansatzes ist dessen gute Interpretier- und Plausibilisierbarkeit, d. h. die Koeffizienten $\beta_{j,i}$ geben den (isolierten) Einfluss des j-ten Objektmerkmals auf den Wert des i-ten Objektes wieder. Handelt es sich z. B. bei $X_{j,i,}$ um das Alter und bei $P_{i,}$ um den Objektpreis des i-ten Objektes in CHF, so gibt $\beta_{j,i}$ die Preissensitivität von Objekt i in Bezug auf das Gebäudealter an (Um wieviel ändert sich der Transaktionspreis bei einer Veränderung des Gebäudealters um ein Jahr?)[5]. Ein Nachteil des hedonischen Ansatzes ist, dass (nicht-lineare) Kreuzbeziehungen zwischen den Objektmerkmalen (z. B. zwischen Objektalter und Objektgrösse) bei einer linearen bzw. log-linearen Modellspezifikation nicht berücksichtigt sind[6].

2.2 Robuste Schätzung des hedonischen Modells

In der einschlägigen Literatur wird üblicherweise unterstellt, dass die Objektpreise P_i im Querschnitt durch eine Lognormal-Verteilung hinreichend genau beschrieben werden können, was jedoch umstritten ist (siehe z. B. Takaaki et al. (2011)). Aus diesem Grund werden die hedonischen Bewertungsmodelle zusätzlich zur OLS-Methode mit einer robusten Methode geschätzt («iterative least squares»), bei der die statistischen Ausreisser weniger gewichtet werden, verglichen mit dem OLS-Verfahren[7]. Im Vergleich zu den OLS-Schätzungen weisen die robusten Parameterschätzungen identische Vorzeichen und eine etwas geringere Signifikanz auf.

2.3 Machine-Learning-Verfahren (ML)

In dieser Studie werden mit dem RF- und GB-Verfahren sowie mit ANN die drei am häufigsten im Zusammenhang mit Regressionsproblemen eingesetzten ML-Verfahren verwendet. Der Vorteil dieser Verfahren im Vergleich zur hedonischen Bewertungsmethode ist, dass weder a priori-Annahmen hinsichtlich nicht-linearer Interaktionen zwischen den Erklärungsvariablen noch bezüglich ihrer Transformationen getroffen werden müssen. Die Beziehungen zwischen den Erklärungsvariablen untereinander und zwischen den Erklärungsvariablen und der zu prognostizierenden Variable werden vielmehr durch die entsprechenden Algorithmen mit Hilfe eines Trainingsdatensatzes «trainiert». Mittels eines Testdatensatzes kann anschliessend die Prognosequalität des optimalen Modells quantifiziert werden. Die Prognosequalität wird üblicherweise mittels des mittleren quadratischen Prognosefehlers («mean squared error») bzw.

[5] Bei einer log-linearen Spezifikation des Bewertungsmodells (d. h. das Alter und der Transaktionspreis gehen in logarithmierter Form in die Bewertungsgleichung ein) kann der Koeffizient $\beta_{j,i}$ approximativ als Elastizität interpretiert werden, d. h. er gibt an, um wieviel Prozent sich der Transaktionspreis bei einer einprozentigen Änderung des Objektalters ändert.

6 Zur OLS-Schätzung der hedonischen Modelle in der Software R wurde die Funktion «lm» benutzt.

7 Zur robusten Schätzung des hedonischen Modells in der Software R wurde die Funktion „rlm“ verwendet.

des mittleren absoluten Prognosefehlers («mean absolute error»), berechnet mit Daten des Testdatensatzes, quantifiziert.

- **Random Forest (RF)**

Das u. a. von Breiman (2001) entwickelte Random-Forest-Verfahren basiert auf mehreren untereinander unkorrelierten Entscheidungsbäumen, mit denen eine bestimmte (quantitative bzw. kategoriale) Zielvariable prognostiziert werden soll. Die Steuerungsparameter («Hyperparameters») dieses Verfahrens, die massgebend sind für den Erklärungsgehalt des optimalen RF-Modells, sind einerseits die maximale Zahl der im Algorithmus berücksichtigten Entscheidungsbäume und andererseits die Zahl der Erklärungsvariablen pro Entscheidungsbaum, d. h. wie viele Merkmale an jedem Knoten des Baums berücksichtigt werden, die als Kriterium für die Aufteilung der Stichprobe dienen[8].

- **Gradient Boosting (GB)**

Das Gradient-Boosting-Verfahren wurde u. a. von J. Friedman (1999) entwickelt und beruht ebenso wie das Random-Forest-Verfahren auf einer vorgegebenen Zahl an Entscheidungsbäumen, die im Gegensatz zum letzteren Verfahren jedoch nicht notwendigerweise unkorreliert sein müssen. Im Vergleich zum Random-Forest-Verfahren ist die Zahl der im Algorithmus pro Entscheidungsbaum berücksichtigen Erklärungsvariablen (maxtree) verhältnismässig klein[9].

- **Künstliche neuronale Netzwerke (ANN)**

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. versteckte Schichten) $g_m(\sum_{j=1}^{k} \omega_j \, log(X_{j,i})) =$

[8] Zur Schätzung des RF-Modelles wird im Statistikpaket R die Funktion «RandomForest» mit den Hyperparametern maxnodes=10 und ntree=500 verwendet. Aus Gründen ungenügender Rechenkapazität konnte keine Optimierung in Bezug auf die Hyperparameter durchgeführt werden.

[9] Zur Schätzung des GB-Modells wird in R die Funktion „gbm“ mit den Parametern *distribution=gaussian“ und n.trees = 1‘000) verwendet.

$g_m(v_j)$ der Immobilie darstellen, wobei die Eigenschaften wiederum nicht-lineare Funktionen der k für die Objektbewertung relevanten Objektmerkmale sind:

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

Die k nicht-linearen Aktivierungsfunktionen g_m werden nicht wie die Parameter ω_j geschätzt, sondern a priori spezifiziert, z. B. Sigmoid $g_m=1/(1+e^{v_j})$, $v_j= \omega_j \, log(X_{j,i})$[10].

3. Empirische Ergebnisse

3.1 Daten

Sämtliche der Studie zugrunde liegenden Daten stammen von der SRED-Datenbank[11] der Periode 1. Quartal 2000 bis 1. Quartal 2020. Bei den Preisen handelt es sich ausschliesslich um Transaktionspreise aus Freihandverkäufen von Objekten (EFH: Einfamilienhäuser; EGTW: Eigentumswohnungen) in der gesamten Schweiz. In den Tabellen 1 und 2 sind die wichtigsten univariaten Statistiken der Verteilungen der Objektpreise und der erhobenen Objektmerkmale aufgeführt.

Tabelle 1: Deskriptive Statistiken für Transaktionspreise und Objektmerkmale, EFH; Quelle: SRED-Database (2000 Q1–2020 Q1), N = 118'345

Variable	Mittelwert	Stand.-Abw.	Median	Min.	Max.	Transfor.
Transaktionspreis (CHF)	904'717	554'843.9	760'000	160'000	6'000'000	log
Kubatur (m^3)*	854.9	307.5	798	250	2'500	log
Grundstückfläche (m^2)	653.3	426	556	100	2'999	log
Transaktionspreis/ Grundstückfläche	1'866	1'435	1'495	57	32'026	log
Zahl Zimmer (ohne Küche, Bad)	5.58	1.36	5	2	12	log
Zahl Nasszellen	2.85		2	1	4	log
Zahl Garagenplätze	-	-	-	0	> 2	log
Gebäudealter	45.5	31	39	0	170	log
Zustand Objekt (1 = schlechteste bis 4 = beste)	2.75	0.89	3	1	4	keine
Qualität Objekt (1 = schlechteste bis 4 = beste)	2.68	0.9	2	1	4	keine

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

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

Qualität Mikrolage (1 = schlechteste bis 4 = beste)	2.85	0.69	3	1	4	keine
Erst-/Zweitdomizil (1: Erstdomizil)	-	-	1 (97 %)	-	-	keine
freistehend/zusammengebaut (1: freistehend)	-	-	1 (62 %			keine
Erklärung: * Kubatur des Objektes als sia- oder GVZ-Norm. Die grösseren Volumina generiert die sia-Norm 416 (wenn überhaupt), dies jedoch in sehr geringem Masse. Abweichungen gibt es i. d. R. nur bei nicht ausgebauten Dachgeschossen, da diese nicht oder nur zum Teil angerechnet werden, gemäss Auskunft der Gebäudeversicherung des Kantons Zürich.						

Tabelle 2: Deskriptive Statistiken für Transaktionspreise und Objektmerkmale, EGTW; Quelle: SRED (2000 Q1–2020 Q1), N = 156'932

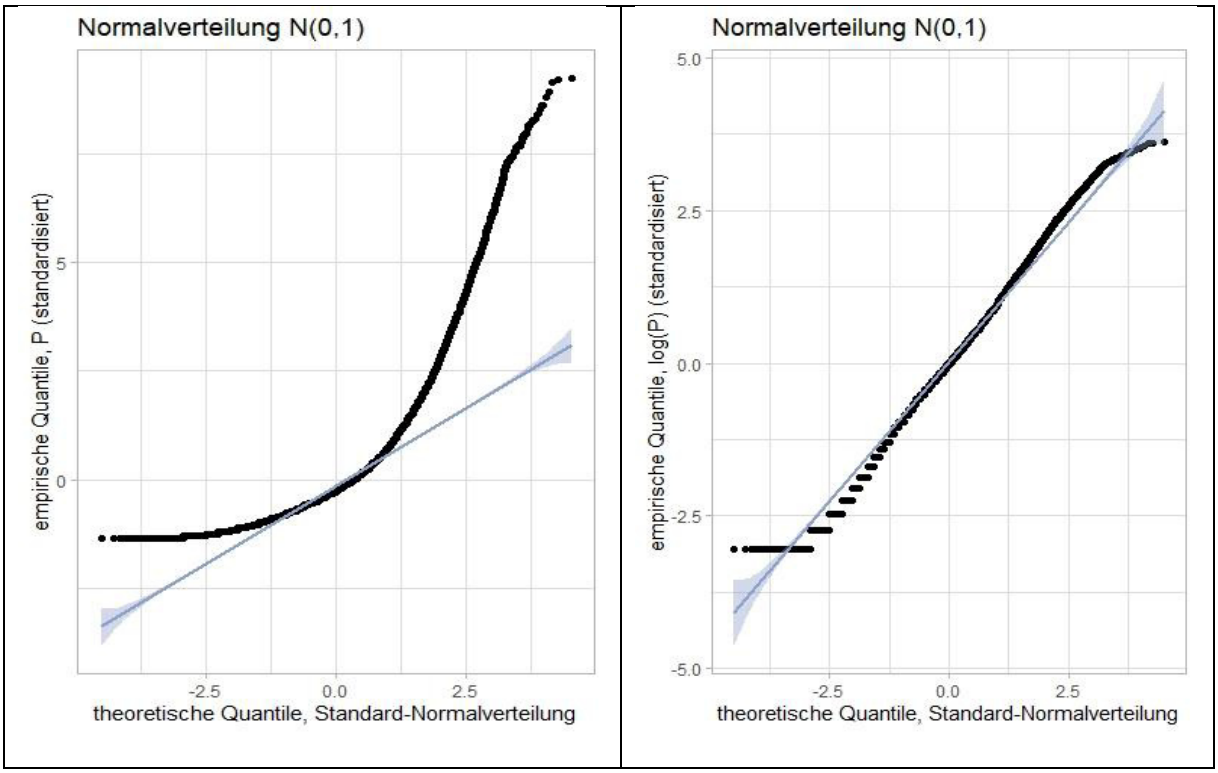
Variable	Mittelwert	Stand.-Abw.	Median	Min.	Max.	Transfor.
Transaktionspreis (CHF)	703'877	454'697	600'000	100'000	4'900'000	log
Nettowohnfläche (m^2)	108.9	36.9	107	30	250	log
Transaktionspreis/ Nettowohnfläche*	6377	2'991	5625	1'500	25'000	-
Zahl Zimmer (ohne Küche, Bad)	3.80	1.11	4.00	1	10	log
Zahl Nasszellen	1.83	0.51	2	1	4	log
Zahl Garagenplätze	-	-	-	0	> 2	log
Objektalter (Jahre)	26.8	22.85	19	0	170	log
Zustand Objekt (1=schlechteste bis 4 = beste)	3.20	0.9	1	1	4	Keine
Qualität Objekt (1 = schlechteste bis 4 = beste)	2.87	0.89	3	1	4	Keine
Qualität Mikrolage (1 = schlechteste bis 4 = beste)	2.81	0.68	3	1	4	Keine
Erst-/Zweitdomizil (1: Erstdomizil)	-	-	1 (90 %)	-	-	Keine

Gemäss den Tabellen 1 bzw. 2 bezieht sich die Median-Transaktion auf ein 1981 bzw. 2001 erstelltes Objekt mit fünf bzw. vier 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. Bei den Einfamilienhäusern wurde ein Preis pro m^2 Grundstücksfläche zwischen CHF 57 (Minimum) und CHF 32'000 (Maximum) verzeichnet, wobei der Medianwert CHF 1'500 betrug.

3.2 Extremwert-Analyse

Für die statistische Modellierung von Immobilienpreisen sind deren Verteilungseigenschaften zentral. Zu ihrer Beschreibung werden die Häufigkeitsverteilung bzw. die Logarithmen der Immobilienpreise mittels einem Quantil-Quantil-Plot[12] (qq-Plot) mit einer Normalverteilung verglichen (siehe die Grafiken 1a und 1b). Je besser die empirische Häufigkeitsverteilung mit der Normalverteilung übereinstimmt, desto eher kann davon ausgegangen werden, dass die Immobilienpreise bzw. deren Logarithmen mit einer Normalverteilung beschrieben werden können.

Abbildung 1a und 1b: qq-Plot P_i bzw. qq-Plot log(P_i) , Eigentumswohnungen (EGTW), N = 156'932



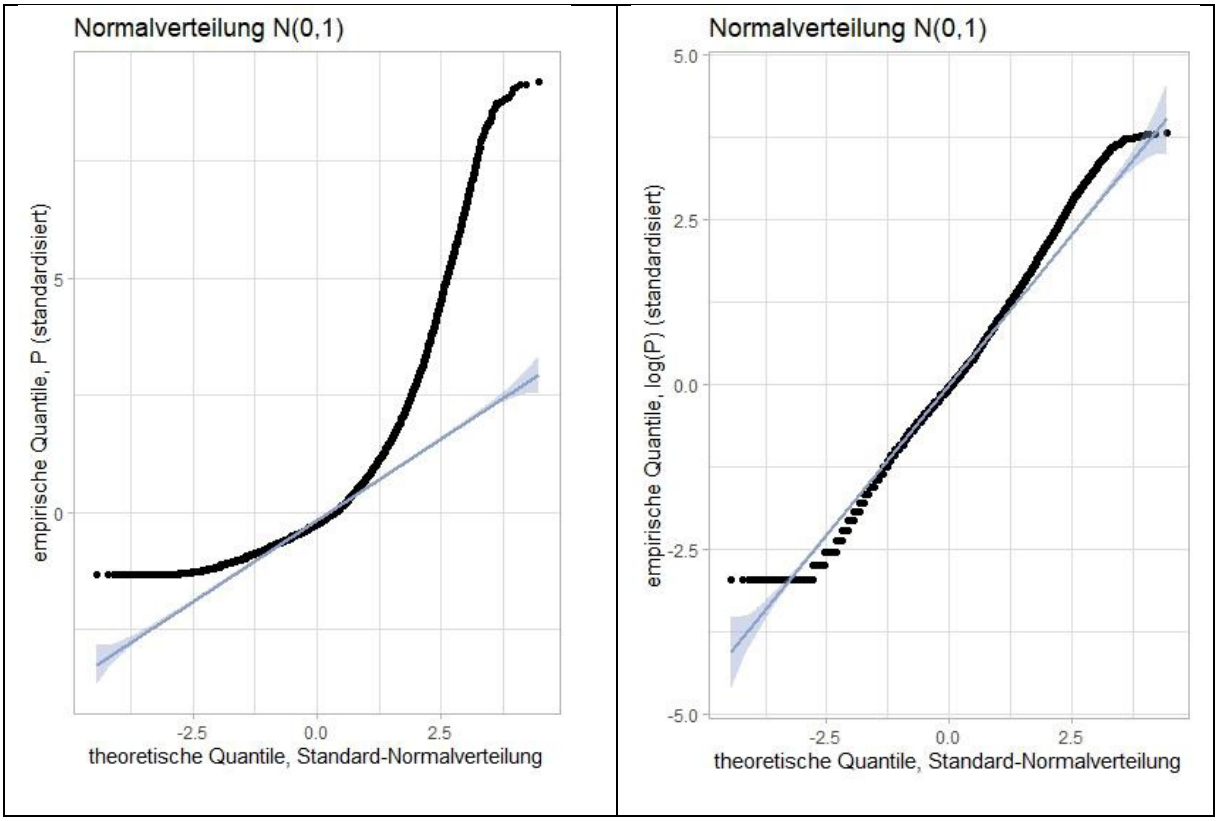
Quelle: SRED, eigene Berechnungen

Wie aus den Grafiken 1a und 1b hervorgeht, weist die Verteilung der Logarithmen der EGTW-Preise eine recht hohe Übereinstimmung mit der Normalverteilung auf. Das Gleiche gilt für die EFH-Preise: Während die nicht-transformierten EFH-Preise relativ stark von der Normalverteilung abweichen, weisen die logarithmierten EFH-Preise eine hohe Übereinstimmung mit der Normalverteilung auf (siehe die Grafiken 2a und 2b). Die Resultate der explorativen qq-Plots können dahingehend zusammengefasst werden, dass sowohl die EGTW- als auch die EFH-

[12] In einem qq-Plot werden die Verteilungsfunktionen zweier Variablen gegeneinander abgetragen, um ihre Verteilungen zu vergleichen.

Preise relativ gut durch eine Lognormalverteilung[13] beschrieben werden können. Dieser Befund wird durch zahlreiche empirische Studien mit Schweizer (siehe z. B. Sconamiglio et al. (2019)) und mit ausländischen Daten (siehe z. B. Razen et al. (2014)) untermauert[14].

Abbildung 2a und 2b: qq-Plot P_i bzw. qq-Plot log(P_i) , Einfamilienhäuser (EFH), N = 118'345



Quelle: SRED, eigene Berechnungen

3.3 Hedonischer Ansatz

Als Referenz für die mittels Machine-Learning-Verfahren ermittelten theoretischen Objektpreise dient der etablierte hedonische Ansatz zur Bestimmung von Immobilienpreisen. Es werden deshalb eine lineare und eine log-lineare Spezifikation eines rudimentären hedonischen Modelles formuliert[15]. Die Koeffizienten der quantitativen Variablen der log-linearen Spezifikation können approximativ als Elastizitäten interpretiert werden. Als räumliche Einheit wird der

[13] Bei der Lognormalverteilung bzw. logarithmischen Normalverteilung einer Zufallsvariablen X ist die mit dem Logarithmus transformierte Zufallsvariable X normalverteilt.

[14] Die Studie von Razen et al. (2014) untersuchte die statistische Verteilung für die Quadratmeterpreise von gut 3'000 österreichischen Einfamilienhäusern (selbstbewohnt) und zeigte, dass diese Verteilung relativ gut mit Hilfe einer Lognormalverteilung beschrieben werden konnte.

[15] Da der hauptsächliche Fokus der vorliegenden Analyse nicht auf der hedonischen Immobilienpreismodellierung liegt, werden bei der Formulierung des hedonischen Modells nur die wichtigsten Immobilienattribute verwendet. Es wird insbesondere darauf verzichtet, die Umgebungsvariablen des Objekts (z. B. Aussenlärm, Nähe zu Stationen

Bezirk, in dem sich das zu bewertende Objekt befindet, gewählt[16]. Um die zeitliche Dimension der Objektpreise (rudimentär) abzubilden, wird das Transaktionsjahr als zusätzliche Erklärungsvariable im hedonischen Modell berücksichtigt[17]. Als Schätzmethode für die lineare bzw. log-lineare Spezifikation wird einerseits die Kleinst-Quadrat-Methode (OLS-Methode) und andererseits eine robuste Schätzmethode («iteratively-reweighted-least-squares»-Methode) verwendet; Letztere wird berücksichtigt, weil sie weniger sensitiv ist in Bezug auf Ausreisser in den Daten[18].

3.4 Hedonische Modellschätzung mittels OLS- bzw. robustem Verfahren

Die mittels OLS- bzw. robustem Verfahren ermittelten Koeffizienten des linearen bzw. log-linearen Modells für Eigentumswohnungen sind in Tabelle 3a dokumentiert. Die Überlegenheit der log-linearen gegenüber der linearen Spezifikation bei der EGTW-Bewertung zeigt sich erstens im grösseren Determinationskoeffizienten der log-linearen Spezifikation in Tabelle 3a (R^2, Trainings-Datensatz: 0.8386 statt 0.7211). Zweitens sind die mittels OLS und robustem Verfahren geschätzten Koeffizienten bei der log-linearen Spezifikation nahezu identisch, während sie bei der linearen Spezifikation deutlich voneinander abweichen und teilweise unterschiedliche Vorzeichen aufweisen. Schliesslich sind die Grössenordnungen der geschätzten Koeffizienten bei der log-linearen Spezifikation weitgehend plausibel (z. B. Abnahme des Objektwertes um 2 %, falls Objektalter um 1 % zunimmt, durchschnittliche jährliche Wertzunahme der Eigentumswohnungen von 4.4 % in der Periode 2000-2020).

Tabelle 3a: Hedonische EGTW-Modellschätzung

Lineares Modell: $P_i = \sum \beta_j X_{i,i,} + u_i$			**log-lineares Modell:** $\log(P_i) = \sum \beta_j \log(X_{i,i}) + u_i$		
Erklärungsvariable $X_{i,i,}$	**OLS**	**Robust**	**Erklärungsvariable** $X_{i,i,}$	**OLS**	**Robust**
Nettowohnfläche	6'795***	5'451**	log (Nettowohnfläche)	0.8766***	0.8747***
Zahl Zimmer	-34'020***	-18'240**	log (Zahl Zimmer)	0.0073*	0.0070**
Zahl Nasszellen	102'100***	77'540***	log (Zahl Nasszellen)	0.1200***	0.1179***
Zahl Garagen	8'714***	19'120**	log (Zahl Garagen)	0.0559***	0.0578**
Alter Objekt	50.7***	-154***	log (Alter Objekt)	-0.0198***	-0.0195***

des öffentlichen Verkehrs, kantonaler bzw. kommunaler Steuersatz, Wohnzone der Grundstücke bei EFH (relevant für Ausnützungsziffer) als Erklärungsvariablen zu berücksichtigen.

[16] Bei schweizweit 142 Bezirken und 118'345 Einfamilienhäusern resultieren im Mittel pro Bezirk gut 833 zu bewertende Objekte. Würde als räumliche Einheit die Gemeinde gewählt, resultierten im Mittel bei schweizweit 2'212 Gemeinden gut 53 Objekte, was keine präzisen Schätzungen mehr erlauben würde.

[17] Wegen Multikollinearität wird darauf verzichtet, einen linearen Zeittrend für die Objektpreise zu spezifizieren.

[18] Die Kleinst-Quadrat-Methode minimiert die Summe der quadrierten Bewertungsfehler (mean squared error). Bei robusten Methoden wird die Summe der mit einer bestimmten Funktion gewichteten Bewertungsfehler minimiert, wobei in dieser Studie der Huber-Schätzer Verwendung findet.

Zustand Objekt (ordinal)	35'360***	34'875**	Zustand Objekt (ordinal)	0.0741***	0.0747**
Ausbau Objekt (ordinal)	65'780***	56'577**	Ausbau Objekt (ordinal)	0.0968***	0.0931**
Jahr Transaktion	32'450***	27'562**	log (Jahr Transaktion)	0.0438***	0.0436**
Qualität Mikrolage (ordi.)	86'490***	67'290***	Qualität Mikrolage (ordi.)	0.1133***	0.1111***
Erst-/Zweitdomizil (no-min.)	121'500***	97'396**	Erst-/Zweitdomizil (no-min.)	0.1746***	0.1728**
142 Bezirke (Referenz: Aarau); maximaler bzw. minimaler Bezirkszu-schlag	Zürich: 499'500 *** Franches-Montagnes: -267'900***	Zürich: 421'000*** Franches-Montagnes: -220'892***	142 Bezirke (Referenz: Aarau; maximaler bzw. minimaler Bezirkszug-schlag	Zürich: .62** Franches-Montagnes: -0.44***	Zürich: .61** Franches-Montagnes: -0.45***
Prognosefehler-Statistiken: Trainingsdaten (N = 116'993)					
R^2	**0.7211**	**-**	**R^2**	**0.8386**	**-**
RMSE	**241'008**	**147'803**	**RMSE**	**0.2034**	**0.2351**
MAE	**160'619**	**453'332**	**MAE**	**0.1336**	**0.1333**
Prognosefehler-Statistiken: Testdaten (N = 18'995)					
RMSE	**242'984**	**225'771**	**RMSE**	**0.2343**	**0.2339**
MAE	**161'638**	**675'485**	**MAE**	**0.1785**	**0.1771**
Innerhalb10% (Testdaten)	**-**	**-**	**Innerhalb10%**	**0.3698**	**0.3724**
Innerhalb20% (Testdaten)	**-**	**-**	**Innerhalb20%**	**0.6521**	**0.6662**
Erklärung: *** bzw. **: p-Wert = 0.01 bzw. p-Wert = 0.05 MAE: Mittelwert der absoluten Preisfehler RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler					

Quelle: eigene Berechnung, SRED-Datenbank

Tabelle 3b: Hedonische EFH-Modellschätzung

Lineares Modell: $P_i = \Sigma \beta_j X_{i,i,} + u_i$			**log-lineares Modell:** $\log(P_i) = \Sigma \beta_j \log(X_{i,i}) + u_i$		
Erklärungsvariable $X_{i,i}$	**OLS**	**Robust**	**Erklärungsvariable**	**OLS**	**Robust**
Kubatur (sia bzw. GVA)	565***	500***	Log (Kubatur)	0.45***	0.47***
Grundstücksfläche	214***	180***	Log (Grundstücksfläche)	0.16***	0.16***
Zahl Zimmer	4'015**	2'178**	Log (Zahl Zimmer)	0.09***	0.09***
Zahl Nasszellen	80'890***	67'419***	Log (Zahl Nasszellen)	0.12***	0.12***
Zahl Garagen	7'625***	7'080***	Log (Zahl Garagen)	0.04***	0.03***
Alter Objekt	-1'381***	-1'579***	Log (Alter Objekt)	-0.10***	-0.09***
Zustand Objekt (ordinal)	43'260***	37'695***	Zustand Objekt (ordinal)	0.04***	0.05***
Ausbau Objekt (ordinal)	84'970***	71'452***	Ausbau Objekt (ordinal)	0.09***	0.08***
Jahr Transaktion	33'250***	27'929***	Log (Jahr Transaktion)	0.03***	0.03***
Qualität Mikrolage (ordi.)	101'800***	71'452***	Qualität Mikrolage (ordi.)	0.12***	0.11***
Erst-/Zweitdomizil (no-min.)	120'600***	73'111***	Erst-/Zweitdomizil (no-min.)	0.08***	0.08***
142 Bezirke (Referenz: Aarau); maximaler bzw.	Zürich: 769'500***	Zürich: 616'883*** Pruntrut:	142 Bezirke (Referenz: Aarau; Max. bzw. min.	Zürich: 0.76** Pruntrut:	Zürich: .74***

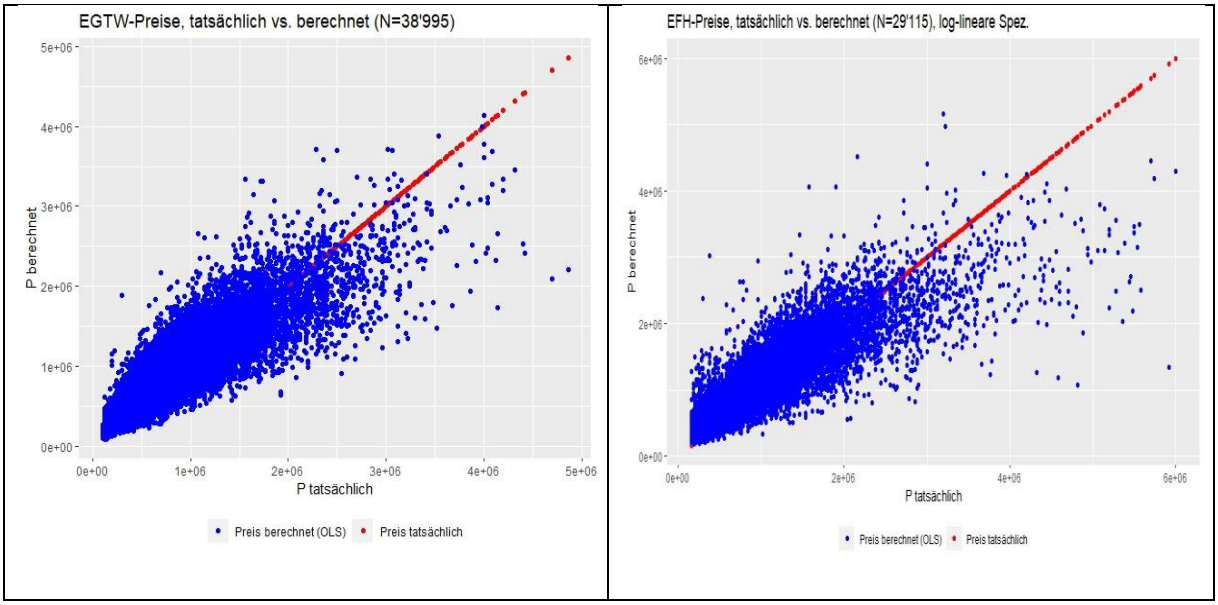
minimaler Bezirkszuschlag	Pruntrut: -392'400***	-366'593***		-0.67***	-0.65***
Prognosefehler-Teststatistiken: Trainingsdaten (N = 87'195)					
R^2 (Trainingsdaten)	**0.6798**	**-**	**R^2 (Training Daten)**	**0.7739**	**-**
RMSE (Trainingsdaten)	**313'781**	**323'577**	**RMSE (Training-Daten)**	**0.2530**	**0.2539**
MAE (Trainingsdaten)	**203'009**	**193'439**	**MAE (Training-Daten)**	**0.1892**	**0.1841**
Prognosefehler-Teststatistiken: Testdaten (N = 29'115)					
RMSE (Testdaten)	**314'625**	**960'630**	**RMSE (Test-Daten)**	**0.2546**	**0.2550**
MAE (Testdaten)	**202'614**	**877'706**	**MAE (Test-Daten)**	**0.1866**	**0.1854**
Innerhalb10% (Testdaten)	**-**	**-**	**Innerhalb10%**	**0.3644**	**0.3701**
Innerhalb20% (Testdaten)	**-**	**-**	**Innerhalb20%**	**0.6469**	**0.6517**
Erklärung:. *** bzw. **: p-Wert = 0.01 bzw. p-Wert=0.05 MAE: Mittelwert der absoluten Preisfehler RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler					

Quelle: eigene Berechnung, SRED-Datenbank

Die Koeffizienten der linearen bzw. log-linearen Spezifikation (1a) bzw. (1b) des hedonischen EFH-Modells sind in Tabelle 3b aufgeführt. Analog zum Modell für Eigentumswohnungen weist die lineare Spezifikation einen um knapp 10 Prozentpunkte tieferen Erklärungsgehalt verglichen mit der log-linearen Spezifikation auf. Bei der log-linearen Spezifikation sind die Koeffizienten des EFH- und des EGTW-Modells einander sehr ähnlich mit Ausnahme des Koeffizienten, der die Altersentwertung des Objektes abbildet (-0.10 versus -0.02). Obwohl die robuste Schätzung der log-linearen Spezifikation eine grössere Streuung der Residuen im Vergleich zur OLS-Schätzung bei Verwendung des Trainingsdatensatzes aufweist, zeichnet sich das erste Verfahren bei Verwendung der log-linearen Spezifikation und der Testdaten durch eine geringfügig kleinere Streuung der Preisfehler aus im Vergleich zu der OLS-Schätzung (0.2542 verglichen mit 0.2562). Auch in Bezug auf den mittleren absoluten Preisfehler (MAE) resultiert bei Verwendung der robusten Schätzung und des Testdatensatzes ein kleinerer MAE im Vergleich zur OLS-Schätzung.

In den Grafiken 4a und 4b werden die mit dem hedonischen Modell berechneten EGTW- und EFH-Preise mit den tatsächlichen Transaktionspreisen unter Verwendung der entsprechenden Testdaten verglichen. Es zeigt sich, dass die berechneten die tatsächlichen Preise ausserhalb der Stützbereichs (d. h. ausserhalb der Trainingsdaten) im Mittel unterschätzen. Weiter ist erkennbar, dass die Streuung der berechneten um die tatsächlichen Preise bei den Eigentumswohnungen kleiner ist als bei den Einfamilienhäusern (siehe Tab. 3a und 3b).

Abbildung 4a und 4b: P_i tatsächlich vs. P_i berechnet (log-lineare Spezifikation), EGTW und EFH, Testdaten



Quelle: eigene Berechnung, SRED-Datenbank

3.5 Implementation von Machine-Learning-Verfahren

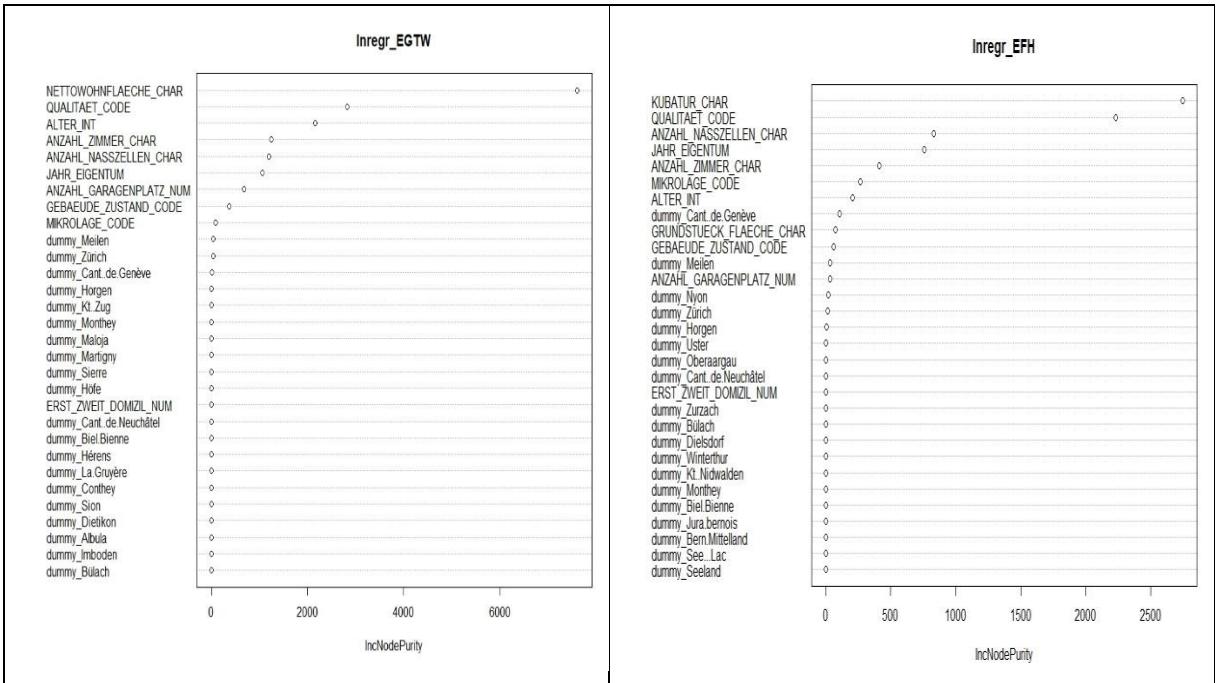
Die Genauigkeit der mittels drei ML-Verfahren ermittelten Objektpreise soll in den nachfolgenden Abschnitten analog zu den hedonischen Modellen unter Verwendung von Trainings- bzw. Testdaten und mit Hilfe von zwei Kennzahlen (RMSE bzw. MAE) quantifiziert werden. Im Gegensatz zur Überprüfung der Preisgenauigkeit der hedonischen Modelle sollen nur die log-lineare Spezifikationen Verwendung finden, da die linearen Spezifikationen des hedonischen Ansatzes einen durchweg tieferen Erklärungsgehalt bei Verwendung der Testdaten aufwiesen. Ein zusätzliches Argument für die log-lineare Modellspezifikation ist, dass die Störprozesse u_i in (1b) approximativ als relative bzw. prozentuale Preisfehler interpretiert werden können.

- **Implementation Random-Forest-Verfahren (RF-Verfahren)**

Das u. a. vom US-Statistiker Breiman (2001) entwickelte Random-Forest-Verfahren zur Lösung von Regressions- bzw. Klassifikationsproblemen basiert auf unkorrelierten Entscheidungsbäumen, mit denen sich Datensätze, bestehend aus metrischen und/oder ordinalen bzw. kardinalen Variablen, mit Hilfe eines Algorithmus möglichst gut «fitten» lassen. In dieser Arbeit wurde das in R implementierte Packet «Random Forest» mit den beiden Eingabeparametern «maximale Zahl der Knoten = 10» (maxnodes) und «ntree = 500» (Zahl der Bäume) gewählt.

Um die Relevanz der einzelnen Objektmerkmale hinsichtlich der Transaktionspreise zu quantifizieren, wurde das Mass «Increase in Node Purity» (IncNodePurity) berücksichtigt. Dieses Mass gibt die Reduktion der Summe der quadrierten Preisfehler wieder, die durch die Berücksichtigung eines bestimmten Objektmerkmals in einem Knoten, verglichen mit der Fehlerquadratsumme, die ohne Berücksichtigung dieses Merkmals generiert wird, eintritt. Die IncNodePurity-Masse sind in Tabelle 4 für die in Bezug auf die für die Eigentumswohnungs- bzw. Einfamilienhauspreise wichtigsten 30 Objektmerkmale wiedergegeben. Zu beachten ist, dass weniger die Absolutwerte der IncNodePurity-Masse als vielmehr deren Verhältnisse relevant sind, d. h. z. B. bei den Eigentumswohnungen ist die Nettowohnfläche mehr als dreimal so relevant in Bezug auf den Transaktionspreis verglichen mit dem Alter des Objektes. Verglichen mit dem anderen in dieser Arbeit berücksichtigten, auf Entscheidungsbäumen basierenden Gradient-Boosting-Verfahren ist die Relevanz der räumlichen Variablen (Mikrolagequalität und vor allem Bezirksdummies) hinsichtlich der Objektpreise deutlich geringer[19].

Tabelle 4: EGTW- bzw. EFH-Preisschätzung mit Random Forest, relative Bedeutung der Erklärungsvariablen («Increase Node Impurity»)



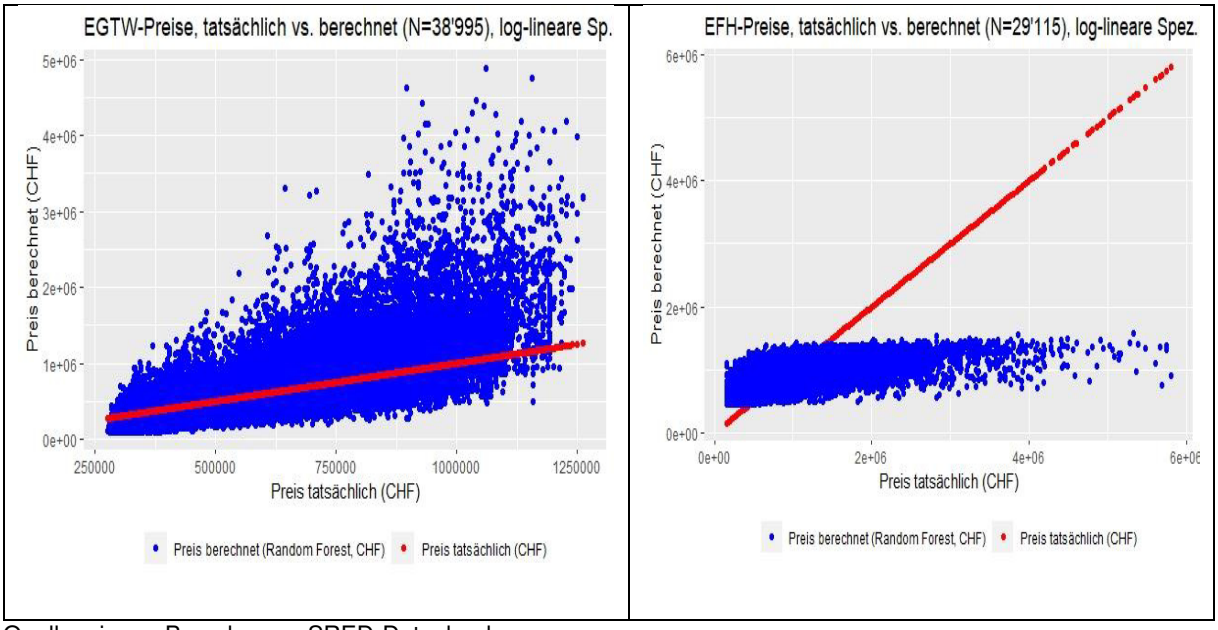
Prognosefehler-Teststatistiken des Random Forest-Verfahrens	**Eigentumswohnungspreise (P_i)**	**Einfamilienhauspreise (P_i)**
RMSE (Trainingsdaten)	**0.3877**	**0.4017**

[19] Diese Tatsache kann auch damit zu tun haben, dass in der eingesetzten Software («R») die räumlichen Bezirksvariablen beim Random-Forest-Verfahren als Dummyvariablen modelliert werden mussten, während diese 143 kardinalen Variablen beim hedonischen Ansatz und dem mittels Gradient-Boosting-Verfahren ermittelten Entscheidungsbäumen als Faktorvariablen spezifiziert werden konnten.

RMSE (Testdaten)	**0.3895**	**0.4066**
MAE (Trainingsdaten)	**0.3035**	**0.3074**
MAE (Testdaten)	**0.3047**	**0.3074**
Innerhalb10% (Testdaten)	**0.2123**	**0.2180**
Innerhalb20% (Testdaten)	**0.4133**	**0.4216**
Erklärung: MAE: Mittelwert aus absoluten Preisfehlern RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler		

Im unteren Teil der Tabelle 4 sind die beiden Kennzahlen (Wurzel aus mittlerer Fehlerquadratsumme, mittlerer absoluter Preisfehler) zur Beurteilung der Preisfehler der mittels des Random-Forest-Verfahrens ermittelten Entscheidungsbäume aufgeführt. Verglichen mit den anderen in dieser Arbeit berücksichtigten Ansätze bzw. Verfahren (hedonischer Ansatz, Gradient-Boosting-Verfahren, Neuronal Network) weist das Random-Forest-Verfahren gemäss beiden Kennzahlen sowohl bei Verwendung der Trainings- als auch der Testdaten die grössten Preisfehler auf.

Abbildung 5a, 5b: P_i tatsächlich vs. P_i berechnet, EGTW u. EFH, Random Forest, Testdaten



Quelle: eigene Berechnung, SRED-Datenbank

In den Grafiken 5a und 5b werden die mittels des Random-Forest-Verfahrens berechneten EGTW- und EFH-Preise mit den tatsächlichen (Transaktions-)Preisen für die EGTW- bzw. EFH-Testdaten verglichen. Es ist erkennbar, dass die berechneten die tatsächlichen Preise ausserhalb des Stützbereichs (d. h. ausserhalb des Trainingsdatensatzes) vor allem bei den Einfamilienhäusern im Mittel unterschätzen. Weiter ist festzuhalten, dass die Streuung der berechneten Preise bei den Eigentumswohnungen grösser ist als bei den Einfamilienhäusern.

- **Implementation Gradient Boosting-Verfahren (GB-Verfahren)**

Das Gradient Boosting ist ein Verfahren, das u. a. von Hastie et al. (2009) zur Lösung von Klassifikations- und Regressionsproblemen entwickelt wurde, bei denen eine quantitative bzw. eine kategoriale zu erklärende Variable möglichst gut mittels einer vorgegebenen Liste von quantitativen und/oder kategorialen Variablen erklärt bzw. prognostiziert wird. Analog zu anderen ML-Verfahren (z. B. Random Forest) basiert die Gradient-Boosting-Methode auf Entscheidungsbäumen, bei denen eine vorgegebene Verlustfunktion (d. h. Summe der quadrierten Abweichungen des theoretischen vom tatsächlichen Transaktionspreis) schrittweise minimiert wird. Im Gegensatz zur hedonischen Methode erlaubt das Gradient-Boosting-Verfahren keine Schätzung von quantitativen (interpretierbaren) Koeffizienten, sondern lässt nur Rückschlüsse hinsichtlich der relativen Bedeutung einzelner Objektmerkmale in Bezug auf den Transaktionspreis eines bestimmten Objektes zu. Diese Relevanz der einzelnen Objektattribute ist in der unten stehenden Tabelle 4a aufgeführt. Zusätzlich enthält die Tabelle die Kennzahlen zur Messung der Preisgenauigkeit.

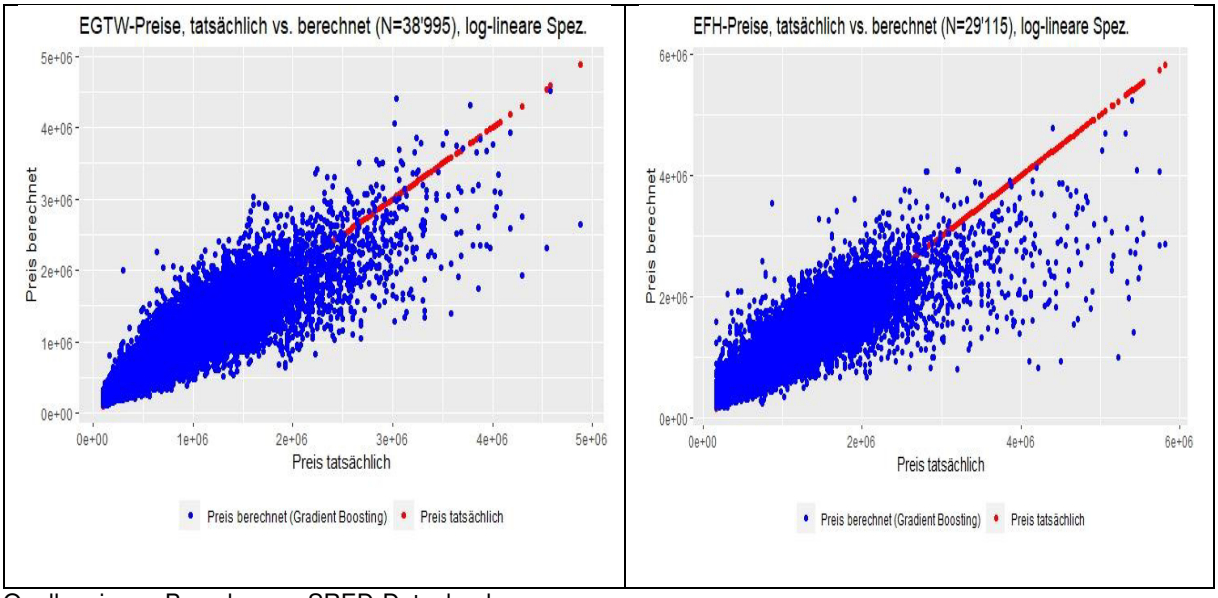
Tabelle 5: EGTW- bzw. EFH-Preisschätzung mit Gradient Boosting, relative Bedeutung der Erklärungsvariablen, Anzahl Entscheidungsbäume = 1'000

Erklärungsvariable	**Eigentumswohnungspreise (P_i)**	**Einfamilienhauspreise (P_i)**
Kubatur	-	0.25
Grundstücksfläche	-	0.03
Nettowohnfläche	0.44	-
142 Bezirke (räumliche Variable)	0.22	0.39
Jahr Transaktion	0.12	0.11
Ausbau Objekt	0.09	0.12
Alter Objekt	0.07	0.04
Qualität Mikrolage	0.03	0.04
Zahl Nasszellen	0.01	0.01
Zahl Garagen	0.01	0.001
Erst-/Zweitdomizil	0.005	0.001
Zustand Objekt	0.003	0.009
Zahl Zimmer	0.002	0.005
Summe	1.0	1.0
Prognosefehler-Teststatistiken des Gradient-Boosting-Verfahrens	**Eigentumswohnungspreise (P_i)**	**Einfamilienhauspreise (P_i)**
RMSE (Trainingsdaten)	**0.2287**	**0.2408**
RMSE (Testdaten)	**0.2303**	**0.2484**
MAE (Trainingsdaten)	**0.1736**	**0.1806**
MAE (Testdaten)	**0.1745**	**0.1813**

Innerhalb10% (Testdaten)	**0.3899**	**0.3774**
Innerhalb20% (Testdaten)	**0.6656**	**0.6559**
Erklärung: MAE: Mittelwert aus absoluten Preisfehlern RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler		

In den Grafiken 6a bzw. 6b sind die tatsächlichen Preise der Testdaten den mit Hilfe des GB-Verfahrens berechneten theoretischen Preisen gegenübergestellt. Sowohl bei den Einfamilienhäusern als auch bei den Eigentumswohnungen scheinen die theoretischen die tatsächlichen Preise des Testdatensatzes zu unterschätzen.

Abbildung 6a, 6b: P_i tatsächlich vs. P_i berechnet, EGTW u. EFH, Gradient Boosting, Testdaten



Quelle: eigene Berechnung, SRED-Datenbank

- **Implementation Künstliches Neuronales Netzwerk (ANN-Verfahren)**

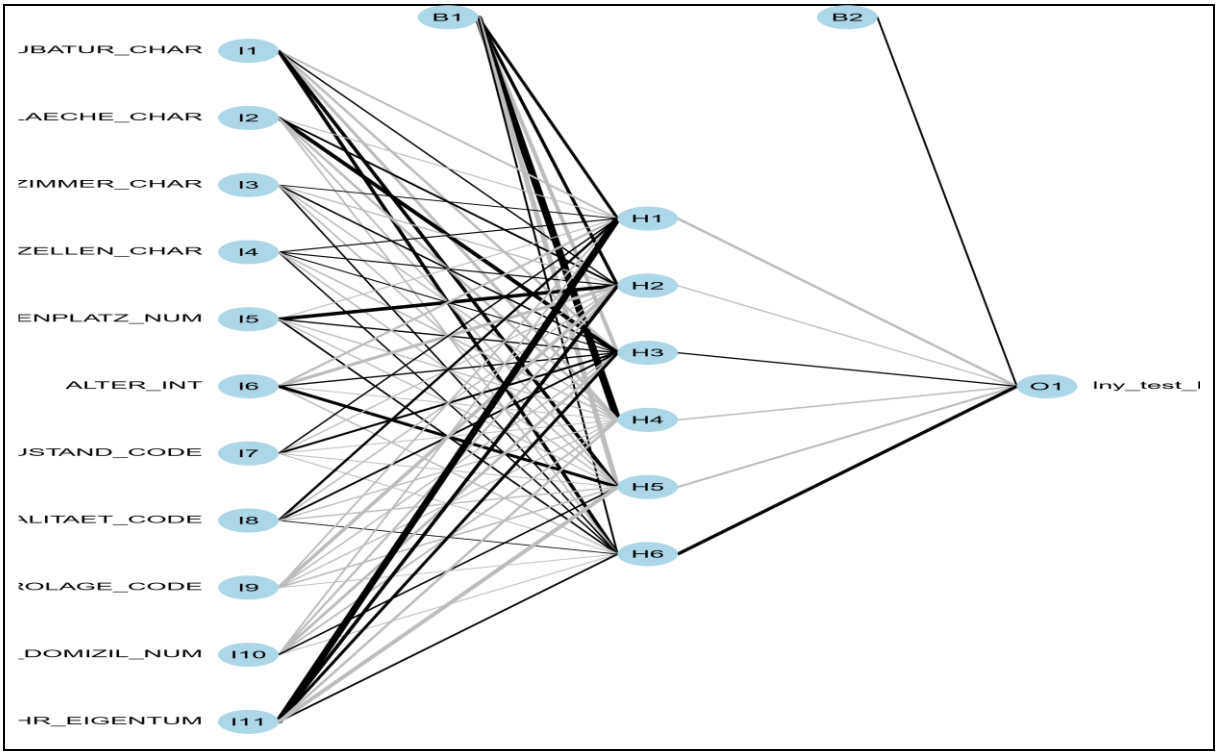
Ein Nachteil des (linearen) hedonischen Ansatzes der Immobilienbewertung ist, dass nicht-lineare Zusammenhänge zwischen den Objektmerkmalen und Objektpreisen nicht abgebildet werden können. Mit neuronalen Netzen lassen sich hingegen nicht-lineare und miteinander verknüpfte Zusammenhänge zwischen Variablen modellieren, wobei die Formen dieser Zusammenhänge nicht explizit formuliert werden müssen. Im Folgenden wird ein neuronales Netz mit einer verborgenen Schicht («Layer») mit sechs Knoten verwendet, um die Preise von Eigentumswohnungen bzw. Einfamilienhäusern mit deren Objekteigenschaften zu erklären (siehe Abbildungen 7a und 7b, ohne die Bezirksdummies). Eine Voraussetzung für die sinn-

volle Anwendung eines ANN im Rahmen eines Regressionsmodells ist, dass die Erklärungsvariablen und die (quantitative) Zielvariable, die normalerweise in verschiedenen Einheiten erfasst werden und unterschiedliche Grössenordnungen aufweisen, standardisiert werden. Dies macht eine Normalisierung der in das ANN eingehenden Daten notwendig, d. h. die normalisierten Daten sollen nur Werte in einem eng begrenzten Bereich annehmen. Für die Objektschätzungen mittels ANN wurden sämtliche quantitative Variablen mit Hilfe folgender Funktion normalisiert bzw. standardisiert:

(3) $x\_norm = (x - x\_min)/(x\_max - x\_min)$,

wobei x_min bzw. x_max das Minimum bzw. das Maximum Variable x ist.

Grafik 7a: EFH-Preisschätzung mit neuronalem Netz mit einer verborgenen Schicht («hidden layer» mit 6 Knoten), Trainingsdaten, ohne Bezirksdummies



Grafik 7b: EGTW-Preisschätzung mit neuronalem Netz mit einer verborgenen Schicht («hidden layer» mit 6 Knoten), ohne Bezirksdummies

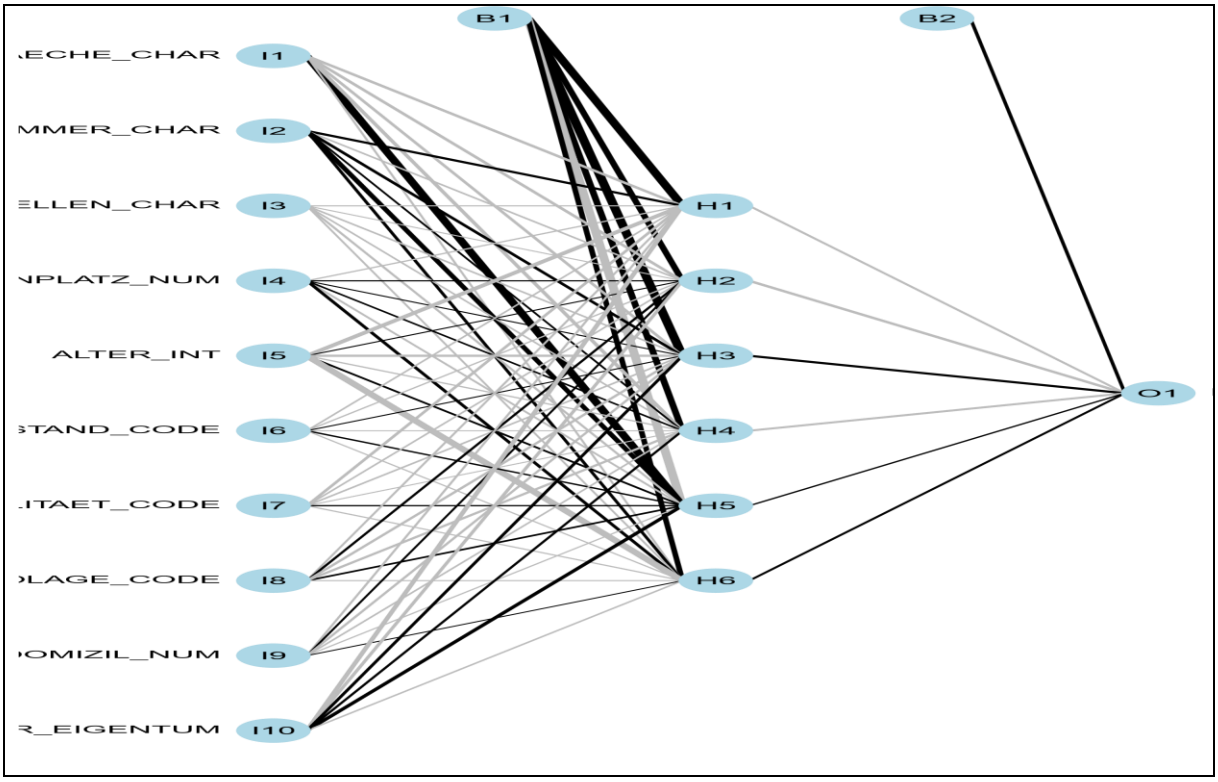
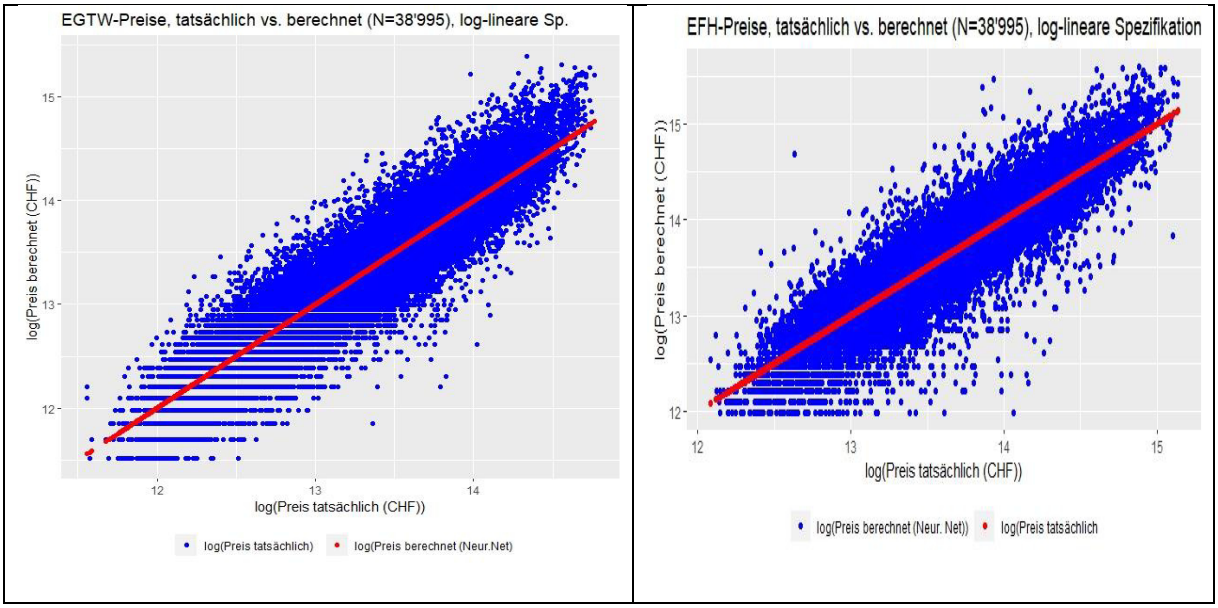


Tabelle 6: Prognosefehler-Teststatistiken des neuronales Netzes mit einer verborgenen Schicht («hidden layer») für Eigentumswohnungen und Einfamilienhäuser

Prognosefehler-Teststatistiken	Eigentumswohnungspreise (P_i)	Einfamilienhauspreise (P_i)
RMSE (Trainingsdaten)	**0.0569**	**0.0639**
RMSE (Testdaten	**0.2227**	**0.2317**
MAE (Trainingsdaten)	**0.0430**	**0.0461**
MAE (Testdaten)	**0.1674**	**0.1672**
Innerhalb10%	**0.4066**	**0.4184**
Innerhalb20%	**0.6995**	**0.7002**
Erklärung: MAE: Mittelwert aus absoluten Preisfehlern RMSE: Wurzel aus Mittelwert der quadratischen Preisfehler		

Beim ANN-Verfahren werden die ω_j- bzw. die g_m-Parameter in Gleichung (2) durch die Minimierung einer Zielfunktion (Summe der quadrierten Abweichungen der theoretischen von den tatsächlichen Transaktionspreisen) iterativ bestimmt. Als Startwerte werden die ω_j- bzw. die g_m-Parameter auf null gesetzt.

Abbildung 7a, 7b: P_i tatsächlich vs. P_i berechnet, EGTW, EFH, Neuronales Netz, Testdaten



4. Zusammenfassung

Voraussetzung für eine sinnvolle Anwendung von ML-Methoden im Bereich Immobilienbewertung ist, dass die Ergebnisse dieser Verfahren auf eine transparente Art mit einer bereits etablierten Methode, die plausibilisierbare Ergebnisse liefert, verglichen werden können (z. B. hedonisches Verfahren und/oder Expertenschätzungen).

4.1 Sind ML-Verfahren generell genauer als hedonische Modelle?

Die Ergebnisse der Literaturanalyse und der Analyse der EFH- bzw. EGTW-Preisfehler von RF, GB und ANN in den Tabellen 4-6 zeigen, dass ML-Verfahren üblicherweise, gemessen an ihren durchschnittlichen Prognosefehlern, den hedonischen Bewertungsansätzen überlegen sind. So ist der mittlere quadratische bzw. mittlere absolute Prognosefehler von GB und ANN sowohl bei EFH als auch bei EGTW teilweise deutlich tiefer als bei der log-linearen Spezifikation des hedonischen Modells[20]. Zudem ist der Anteil der theoretischen Preise, deren approximative relative Abweichung (Absolutbetrag) vom tatsächlichen Objektpreis kleiner ist als 20 bzw. 10 %, bei ANN grösser als beim hedonischen Ansatz (bei EFH und EGTW). Das Beispiel RF zeigt jedoch, dass ein ML-Algorithmus nicht notwendigerweise eine verglichen mit einem «optimal» spezifizierten hedonischen Modell überlegene Preisqualität und damit einen

[20] Die Wurzel aus dem mittleren quadrierten prozentualen Prognosefehler des hedonischen Modells (log-lineare Spezifikation) für Eigentumswohnungen beträgt 26.18 %, während dieselbe Grösse unter Anwendung eines neuronalen Netzwerkes mit einer verdeckten Schicht und sechs Knoten 24.11 % beträgt, wobei beide Verfahren von einem identischen Datensatz ausgehen.

signifikant kleineren (durchschnittlichen) Preisfehler garantiert. Dieses Argument gilt umso mehr, als auch ML-Algorithmen mittels der notwendigen «Hyperparameters» optimal eingestellt werden müssen.

Tabelle 7: Prognosefehler-Teststatistiken, hedonisches Modell vs. drei ML-Verfahren, Testdaten

	RMSE		MAE		Innerhalb10%		Innerhalb20%	
	EGTW	EFH	EGTW	EFH	EGTW	EFH	EGTW	EFH
Hedonisches Modell (OLS)	0.2343	0.2550	0.1785	0.1854	0.3698	0.3701	0.6521	0.6517
Random-Forest-Verfahren	0.3895	0.4017	0.3047	0.3074	0.2123	0.2180	0.4133	0.4216
Gradient-Boosting-Verfahren	0.2303	0.2484	0.1745	0.1813	0.3899	0.3774	0.6656	0.6559
Neuronales Netzwerk-Verf.	**0.2227**	**0.2317**	**0.1674**	**0.1672**	**0.4066**	**0.4184**	**0.6955**	**0.7072**

Quelle: eigene Berechnung, SRED-Datenbank

Bemerkenswert an den mittels der Anwendung von ANN erhaltenen theoretischen Objektpreisen für Einfamilienhäuser und ihre in Tabelle 6 dokumentierten Preisfehler ist die Tatsache, dass sie, verglichen mit den in Sconamiglio et al. (2019) aufgeführten Preisfehlern seiner hedonischen Modellspezifikation, praktisch gleichauf liegen, d. h. nahezu identische durchschnittliche Preisfehler aufweisen. Dies ist umso erstaunlicher, als das in Tabelle 6 dokumentierte EFH-Modell mit nur zwölf Modellvariablen auskommt, während die Spezifikation von Sconamiglio et al. (2019) eine ungleich grössere Zahl an Modellvariablen benötigt (insgesamt 34, darunter 13 Objekt- und 21 Umgebungsvariablen).

4.2 Abschliessende Wertung hedonischer Modelle vs. ML-Verfahren

Ob die ML-Verfahren die traditionellen hedonischen Ansätze der Immobilienbewertung zu ersetzen imstande sind, ist auch gemäss den in dieser Studie aufgeführten Resultaten (noch) eine offene Frage. Hinsichtlich der Prognosequalität erweisen sich die ML-Verfahren zwar überwiegend den hedonischen Modellen als überlegen, doch in Bezug auf die intuitive Verständlichkeit und Kommunizierbarkeit der Modelle sind die ML-Verfahren den hedonischen Ansätzen deutlich unterlegen. Die Qualität der mittels ML-Verfahren berechneten Modelle hängt einzig von ihrer Prognoseleistung, d. h. von den den Modellen zugrundeliegenden Algorithmen ab. Immobilienspezialisten und -spezialistinnen ohne zusätzliche Kenntnisse in ML haben ausser der Prognoseleistung keinerlei Möglichkeit, die mittels ML-Verfahren berechneten theoretischen Preise zu plausibilisieren[21].

[21] Die Kommunikation der hedonischen Objektpreise an nicht mit hedonischen Modellen vertraute Immobilienspezialisten und -spezialistinnen ist an sich schon nicht einfach. Dementsprechend schwieriger dürfte die Kommunikation von mittels ML-Methoden berechneten Objektpreise an nicht mit ML-Algorithmen vertraute Immobilienspezialisten und -spezialistinnen sein.

5. 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 gelten die analogen Definitionen, wobei MAE bzw. RMSE als mittlere absolute bzw. Wurzel aus der mittleren quadratischen (relativen) Preisabweichungen interpretiert werden können[22].

InnerhalbXX% gibt den Anteil der berechneten Preise an der Stichprobe an, der eine maximale (relative) 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.

[22] $\log(P_i^{berechnet}) - \log(P_i^{tatsächlich})$ 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: $\log(1+x) \approx x$.

6. Literaturverzeichnis

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Friedman, J. (1999). *Stochastic gradient boosting. Technical report.* Standford: Standford University.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statstical learning: Data mining, inference, and predictions.* Standford: Springer Series in Statistics.

Kok, N. K.-L., & Martinez-Barbose, C. A. (2017). Big data in real estate? From appraisal to automated valuation. *The Journal of Portfolio Management, Special Real Estate Issue*, 202-211.

Razen, A., Brrunauer, W., Klein, N., Kneib, T., Lang, S., & Umlauf, N. (2014). *Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression.* Insbruck: University of Innsbruck, Working Papers in Economics and Statistics.

Sconamiglio, D. M., Bourassa, S., & Hoesli, M. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Reserarch, 12*(1), 134-150.

Sirmans, G., & Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature, 13*(1), 3-43.

Takaaki, O., Takayuki, M., Chihiro, S., & Tsutomu, W. (2011). *The evolution of house price distribution.* Tokyo: RIETI Discussion Paper Series 11-E-019.