Generieren Machine-Learning-Verfahren präzisere Wohnimmobilienpreis-Prognosen als hedonische Modelle?

Rudolf Marty, Swiss Real Estate Institute, Zürich, 16. Juni 2021

Dieser Beitrag vergleicht die Prognosegüte eines log-linear spezifizierten hedonischen Modells mit der Treffsicherheit der Preisprognosen von drei Machine-Learning-Verfahren unter Verwendung von Schweizer Wohneigentumspreisen.

Machine-Learning-Verfahren in der Immobilienbewertung

Die hedonische Bewertungsmethode gewann in den letzten Jahren vor allem bei der Wertbestimmung von Wohnimmobilien an Bedeutung. Ausgangspunkt ist dabei eine log-lineare Modellspezifikation, die die Logarithmen der Immobilienpreise als Funktion ihrer wertrelevanten logarithmierten Objektmerkmale und ausgewählter Umgebungsvariablen beschreibt. Mittels der Kleinstquadrat-Methode (OLS) werden anschliessend die Modellparameter geschätzt. Mit diesen kann dann der theoretische Preis für ein beliebiges Objekt berechnet werden, für das die preisrelevanten Objektmerkmale bekannt sind (Sirmans et al. 2005). Mit der vermehrten Verbreitung des Maschinellen Lernens (ML) und der Verfügbarkeit wachsender Datenbestände wurden zusätzlich ML-Verfahren zur Immobilienbewertung eingesetzt, z.B. neuronale Netze (ANN, Hestie et al. 2008) und Random Forest (RF, Breimann 2001) sowie Gradient Boosting (GB, Friedman 2000). Das Ziel dieser Studie ist abzuschätzen, ob und in welchem Mass sich die Prognosequalität von Wohneigentumspreisen durch den Einsatz der obigen drei Verfahren des ML im Vergleich zum hedonischen Ansatz verbessern lässt.

In der angelsächsischen Literatur (z.B. Kok et al. 2017) existieren bereits zahlreiche Studien, die die Preisfehler von hedonischen Modellen mit denjenigen von ausgewählten ML-Verfahren u.a. mit Daten des US-Immobilienmarktes vergleichen. Scognamiglio et al. (2019) stellen die Preisfehler von hedonischen Objektpreisen denjenigen von Modellpreisen gegenüber, die mittels drei ML-Verfahren (ANN, RF, GB) für 123'000 Schweizer Einfamilienhäuser berechnet werden. Ihr Befund lautet, dass das GB-Verfahren unter den insgesamt sechs verwendeten Schätzmethoden (OLS-Methode, robuste Schätzung, Mixed-Effect-Verfahren, ANN, GB, RF) gemäss fünf von sechs Statistiken die kleinsten Prognosefehler generiert.

Vergleich eines hedonischen und dreier Machine-Learning-Schätzungsmodelle

Die Daten der vorliegenden Studie stammen von der SRED-Datenbank (1. Quartal 2000–1. Quartal 2020). Bei den Preisen handelt es sich um Transaktionspreise aus Freihandverkäufen von Objekten (EFH: Einfamilienhäuser; EGTW: Eigentumswohnungen) in der gesamten Schweiz. Die Immobilienattribute sind 12 (bei EFH) bzw. 11 (bei EGTW) quantitative und ordinale bzw. kategoriale Objektmerkmale (keine Umgebungsvariablen). Da der Fokus dieser Analyse nicht auf der hedonischen Immobilienpreis-Modellierung liegt, werden bei der Formulierung des hedonischen Modelles nur diejenigen Immobilienattribute verwendet, die in der SRED-Datenbank seit Beginn der Erhebung verfügbar sind.

Die mit OLS-Verfahren geschätzten Koeffizienten des log-linearen Modelles sind in der Tabelle 1 aufgeführt. Die Vorzeichen und Grössenordnungen der Koeffizienten für die EFH bzw. EGTW sind plausibel und fast identisch mit Ausnahme des Koeffizienten für die Anzahl Zimmer und desjenigen, der den Aufschlag für ein Zweit-Domizil wiedergibt. Die Statistiken der Prognosefehler für EGTW weisen verglichen mit denjenigen für EFH eine überlegene bzw. mindestens gleichwertige Prognosegüte auf.

Tabelle 1: Hedonische EFH- bzw. EGTW-Schätzung, Trainingsdaten

log-lineares EFH-Modell: $\log(P_i)=\Sigma \beta_j \log(X_{i,j})+u_i$		**log-lineares EGTW-Modell:** $\log(P_i)=\Sigma \beta_j \log(X_{i,j})+u_i$	
Erklärungsvariable $X_{i,j}$	**Geschätzter Koeffizient**	**Erklärungsvariable $X_{i,j}$**	**Geschätzter Koeffizient**
log(Nettowohnfläche)	-	log(Nettowohnfläche)	0.88***
log(Kubatur)	0.45***	-	-
log(Grundstückfläche)	0.16***	-	-
log(Zahl Zimmer)	0.09***	log(Zahl Zimmer)	0.01*
log(Zahl Nasszellen)	0.12***	log(Zahl Nasszellen)	0.12***
log(Zahl Garagen)	0.04***	log(Zahl Garagen)	0.06***
log(Alter Objekt)	-0.10***	log(Alter Objekt)	-0.02***
Zustand Objekt	0.04***	Zustand Objekt	0.07***
Ausbau Objekt	0.09***	Ausbau Objekt	0.10***
Jahr Transaktion	0.03***	Jahr Transaktion	0.04**
Qualität Mikrolage	0.12***	Qualität Mikrolage	0.11***
Zweit Domizil	0.08***	Zweit Domizil	0.17***
142 Bezirke (Referenz: Aarau); maximaler Bezirkszuschlag bzw. -abschlag	Max. Aufschlag: 0.76** Max. Abschlag:-0.67***	142 Bezirke (Referenz: Aarau; maximaler Bezirkszuschlag bzw. -abschlag	Max. Aufschlag: 0.62** Max. Abschlag:-0.44***
Prognosefehler unter Verwendung des Testdatensatzes; N_{EFH}= 29'115; N_{EGTW} = 38'995			
RMSE	**0.2546**	**RMSE**	**0.2343**
MAE	**0.1866**	**MAE**	**0.1785**
Innerhalb10%	**0.3644**	**Innerhalb10%**	**0.3698**
Innerhalb20%	**0.6469**	**Innerhalb20%**	**0.6521**
Erklärung: *** bzw. **: p-Wert = 0.01 bzw. p-Wert=0.05 **MAE**: Mittelwert der absoluten Preisfehler **RMSE**: Wurzel aus Mittelwert der quadrierten Preisfehler **Innerhalb10%**: Anteil der theoretischen Preise an Stichprobe, die eine maximale (absolute) Abweichung vom Transaktionspreis von 10% aufweisen. (N_{EGTW}=116'993 bzw. N_{EFH} = 87'195)			

Quelle: Eigene Berechnung, SRED-Datenbank

In dieser Studie werden mit dem Random-Forest- und Gradient-Boosting-Verfahren sowie mit dem Verfahren der künstlichen Neuronalen Netzwerke die drei populärsten ML-Algorithmen eingesetzt. Für die Implementation werden die entsprechenden R-Software-Pakete verwendet («randomForest», «gbm», «nnet»). Bei den drei ML-Verfahren werden dieselben Immobilienattribute wie beim hedonischen Modell eingesetzt.

ANN- und GB-Modellpreise liegen näher bei Transaktionspreisen als hedonische Preise

Die Optimalität der theoretischen Preise der vier untersuchten Bewertungsverfahren wird mit Hilfe einer Mincer-Zarnowitz-Regression unter Verwendung des Testdatensatzes überprüft:

$$\log(P_i) = a + b*\log(P^{th}_i) + \varepsilon_i \quad ,$$

wobei P_i der Transaktionspreis und P^{th}_i der theoretische Preis des i-ten Objektes ist und ε_i den (approximativen prozentualen) Preisfehler wiedergibt. Weiter wird die Prognosegüte quantifiziert, indem der Anteil der berechneten Preise an der Stichprobe (Testdatensatz) berechnet wird, der eine maximale absolute Abweichung vom Transaktionspreis von 10% bzw. 20% aufweist.

Die in den Tabellen 2a und 2b aufgeführten Teststatistiken zeigen, dass es sich bei den hedonischen Modellpreisen um optimale Prognosen handelt, die keine systematischen Verzerrungen aufweisen. Allerdings weisen die mittels GB und ANN berechneten Preise einen höheren Erklärungsgehalt in Bezug auf die Transaktionspreise auf und die Anteile der geschätzten Preise, die eine maximale Abweichung von 10% bzw. 20% vom Transaktionspreis aufweisen, liegen 3 bis 4 Prozentpunkte höher als jene des hedonischen Modells.

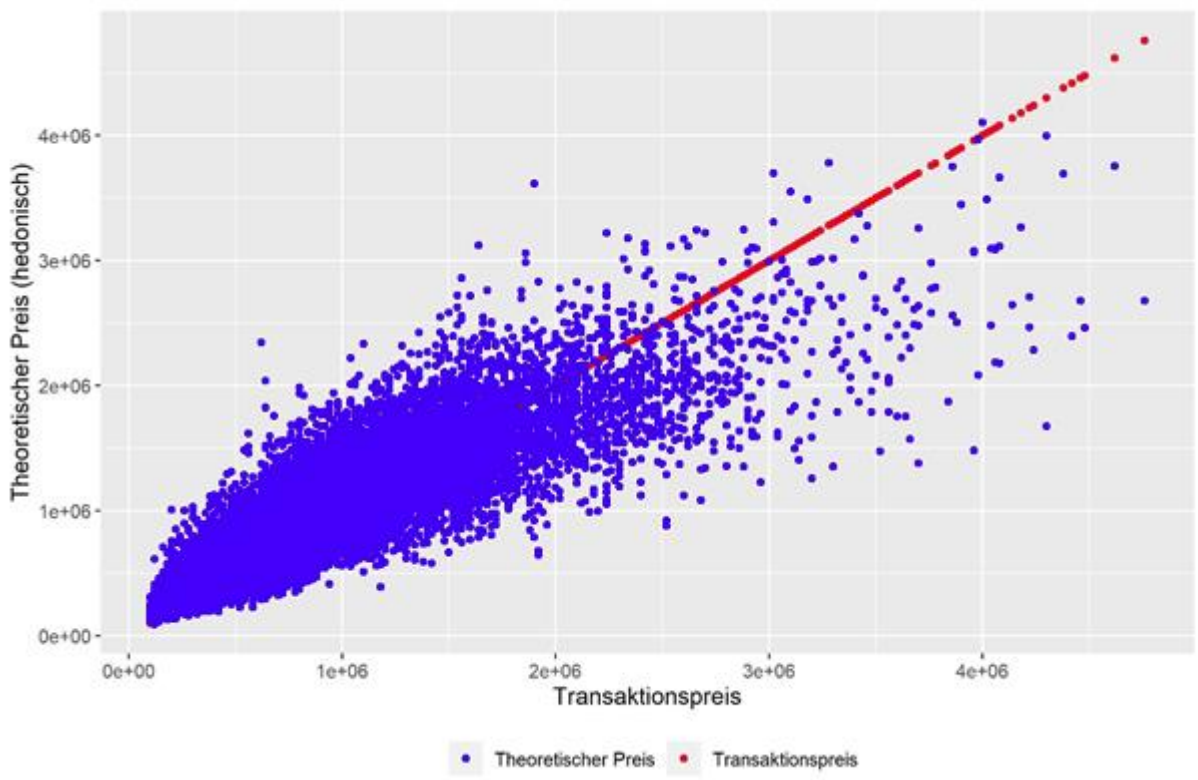
Tabelle 2a: Regression des Transaktionspreises auf theoretischen Preis: EFH, Testdatensatz

	Log-lineares hedonisches Modell	**ML: Random Forest**	**ML: Gradient Boosting**	**ML: künstliches neuronales Netzwerk**
a	-0.04295	**-6.66*****	**-0.1454*****	**-0.1822*****
b	1.00319	**1.49*****	**1.0109*****	**1.0146*****
N	29'115	29'115	29'115	29'115
R^2	0.7725	0.4652	0.7797	0.8108
Innerhalb10%	0.3644	0.2223	0.3794	0.4141
Innerhalb20%	0.6469	0.4245	0.6623	0.6957
Erklärung: *** bzw. **: p-Wert = 0.01 bzw. p-Wert=0.05 (Null-Hypothese: a=0, b=1)				

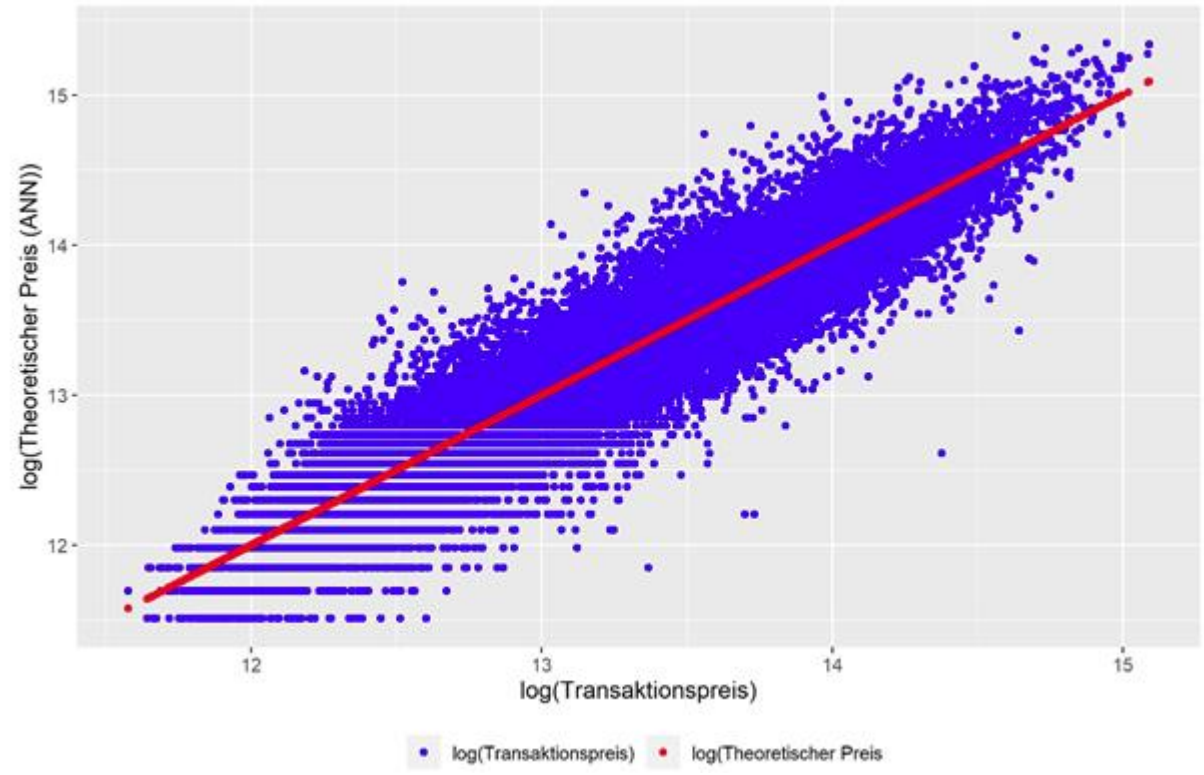
Quelle: Eigene Berechnung, SRED-Datenbank

In den Grafiken 1a bzw. 1b sind die Transaktionspreise der EGTW bzw. deren Logarithmen den theoretischen EGTW-Preisen (hedonische Preise bzw. ANN-Preise für die Testdaten (d.h. ausserhalb des Schätzbereichs des Modelles) gegenüberstellt.

Grafik 1a: Transaktionspreis vs. hedonische Preisschätzung



Grafik 1b: log(Transaktionspreis) vs. log(ANN-Preisschätzung)



Quelle: Eigene Berechnung, SRED-Datenbank

Die Analyse der Abweichungen der Transaktionspreise von theoretischen Immobilienpreisen, die mittels dreier ML-Verfahren berechnet werden, zeigen, dass zwei der drei untersuchten ML-Verfahren aufgrund ihrer Prognosefehler dem hedonischen Ansatz sowohl bei EFH als auch bei EGTW überlegen sind. So ist der Anteil der theoretischen Preise, deren approximative relative Abweichung vom tatsächlichen Objektpreis kleiner ist als 20% bzw. 10%, bei ANN und GB durchwegs grösser als beim hedonischen Ansatz. Diese Ergebnisse bestätigen damit weitgehend die Resultate von Scognamiglio et al. (2019), die sie für EFH erhalten haben, auch für Eigentumswohnungen.

Literatur

Breiman, L. (2001). Random Forests. *Machine Learning,45*, S. 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*, S. 202-211.

Marty, R. (2021). *Machine Learning-Verfahren versus hedonische Bewertung von Wohnimmobilien.* Zürich: HWZ Working Paper Series, No. 1.

Scognamiglio, D. M., Bourassa, S., & Hoesli, M. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Reserarch Vol. 12 No. 1*, S. 134-150.

Sirmans, G., & Macpherson, D. u. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature, Vol. 13 No. 1*, S. S3-43.