FOM Hochschule für Oekonomie & Management

Hochschulzentrum Düsseldorf

Exposé

zur Master-Thesis im Studiengang Big Data / Business Analytics

über das Thema

Analyse des Einflusses von traditionellen sowie neuartigen Faktoren auf die Prognose von Mietpreisen der Stadt Essen

von

Dominic Hernes

Betreuer :	Prof. Dr. Frank Lehrbass
Matrikelnummer :	520014
Abgabedatum :	4. Juli 2021

1 Problemstellung

In der Vergangenheit wurde der Wert einer Immobilie durch eine Immobilienfirma sowie die Bereitschaft eine bestimmte Miete für eine Immobilie durch eine Privatperson zu zahlen an den direkten Gegebenheiten der Immobilie selbst und deren unmittelbarer Umgebung gemessen. Die Lage einer Immobilie war das höchste Gut bei dem Bemessen einer Miete. Jedoch hat sich dieses Verhalten in den letzten Jahren verändert, durch das hohe Angebot von etwaigen Bewertungsportalen und Kartenwerkzeugen im Internet, wie beispielsweise Google, Yelp oder Tripadvisor, welche eine neue Sicht auf die makroskopische Betrachtung einer Immobilie ermöglichen. Mit diesen Werkzeugen lässt sich heutzutage die Lebensqualität messen, welche im Zusammenspiel mit der Lage dem potenziellen Mieter etliche neue Variablen für seine Betrachtung mitgeben. So kann dieser den Grad der Bereitschaft zum Zahlen einer gewissen Miete konkreter bilden und durch den erweiterten Blick auf die frei zugängliche makroskopische Ebene neu entwickeln. Zu dem kann sich der Wert der Immobilie je nach Zielgruppe sowie Quartiersentwicklung unter diesen Variablen sehr differenziert weiterentwickeln. Diesen Einfluss der neuartigen Variablen, welche auch als nicht-traditionellen Kennwerte bezeichnet werden, wollen wir in dieser Arbeit untersuchen.

2 Fragestellung und Zielsetzung

Die zentrale Fragestellung, welche im Fokus dieser angestrebten Masterthesis steht, behandelt inwieweit das Hinzufügen von neuartigen immobilienwirtschaftlichen Kennwerten zu den traditionellen Faktoren die globale Modellgüte der Prognosen verändert und welche Faktoren maßgeblich dafür verantwortlich sind.

Um diese Hypothesen im Laufe der Thesis zu überprüfen, wird der Einfluss von traditionellen sowie den neuartigen Daten auf die Kaltmieten (in €/m²) von Mietobjekten der Stadt Essen analysiert. Dazu wird im ersten Schritt ein Datensatz bestehend aus traditionellen sowie neuartigen Faktoren gebildet, bereinigt und in einer explorativen Analyse näher betrachtet. Auf Basis des zuvor gebildeten Datensatzes wird mithilfe vierer Regressionsmodelle aus den Bereichen des Machine- sowie Deep Learning die obige Fragestellung untersucht werden und anhand einer Stichprobe, welche zuvor entnommen worden ist, angewendet. Die Erkenntnisse aus der quantitativen Untersuchung über den Einfluss, der angesprochenen neuartigen Faktoren, welche auf die Zielvariable abgeleitet wird, sowie die Prognose und die Gütekennwerte aus den Regressionsmodellen werden im finalen Teil dieser Studie vertieft sowie diskutiert.

Im Verlauf der Arbeit sollen weitere Aspekte beleuchtet werden, um das Thema abzurunden:

- Wie betrachtet die Kommune den aktuellen Immobilienmarkt und dessen Entwicklung?
- Was sind die inhaltlichen Unterschiede zwischen traditionellen und neuartigen immobilienwirtschaftlichen Kennwerten?
- Wie valide sind die Prognosen mit Bezug auf die Benchmarks?
- Kann der Grad der Sensitivität neuartiger Faktoren, die der traditionellen Faktoren übertreffen, wenn deren Dimensionen normalisiert werden?
- Können die neuartigen Faktoren allein die gleiche Modellgüte der traditionellen Faktoren erreichen?

3 Methodik

Der Aufbau dieser Thesis erfolgt nach Vorbild des Vertiefungsmodells nach Phillip Mayring, welcher Bestandteile der quantitativen und qualitativen Analysen zu einem kombinierten Gesamtmodell vereint[1]. Auf Basis des Vertiefungsmodells wird die quantitative Untersuchung in Anlehnung an dem CRISP-DM durchgeführt, einem Standartprozessmodell zur Modellierung von Data Mining Projekten im Bereich der computerunterstützen Analyse[2]. Die aus dieser Untersuchung ermittelten Ergebnisse werden daraufhin vertieft, verglichen sowie diskutiert und bilden die qualitative Untersuchung ab.

Im ersten Teil dieser Thesis wird der aktuelle Stand der kommunalen Stadtentwicklung sowie der Entwicklung des Wohnungsbedarfs der Stadt Essen dargestellt[3]. Zudem werden die traditionellen und neuartigen Faktoren definiert und deren Unterschiede beschrieben, welche für die Beantwortung der zentralen Fragestellung mitunter untersucht werden[4].

Die Beschreibung, Aufbereitung und Zusammenführung der Datensätze, bestehend aus den verschiedenen Faktoren der traditionellen und neuartigen immobilienwirtschaftlichen Faktoren erfolgt im zweiten Teil. Der Hauptschwerpunkt dieses Arbeitsschrittes, liegt neben der Beschreibung aller genutzten Datensätze, in der explorativen Datenanalyse sowie

[1] Vgl. *Mayring*, *P.*, 2001.
[2] Vgl. *Wirth*, *R.*, 2000.
[3] Vgl. *Stadt Essen*, *InWIS Forschung*, 2018.
[4] Vgl. *Asaftei*, *G. M.* et al., 2018.

deren Aufarbeitung für den Einsatz in den Prognosemodellen. Die Aufbereitung umfasst verschiedene Methoden von der Datenbereinigung[5], bis hin zur Transformationsmethoden[6] sowie Encoding Varianten[7], um diese bestmöglich für die Analyse vorzubereiten. Im letzten Abschnitt des zweiten Teils wird der gesamte Datensatz für das Training in den Modellen vorbereitet. Dieser wird für diesen Zweck in drei Teile aufgeteilt. Einem Trainings- sowie Testdatensatz, welche für den Einsatz an den Modellen genutzt werden wird und einem zufällig auserwählten Validierungsdatensatz.

Im letzten Teil der Analyse werden nunmehr Prognosemodelle auf Basis der zuvor aufgearbeiteten Daten eingesetzt und durchgeführt, indem diese mithilfe des Training- sowie Testdatensatzes trainiert werden, so dass diese an dem Validierungsdatensatz angewendet werden können. Die Optimierung der Hyperparameter, der genutzten Modelle erfolgt mit Hilfe der Random Search Methode[8]. Die Prognosen des Validierungsdatensatzes bilden dadurch einen essenziellen Teil der Diskussion ab.

Folgende Prognosemodelle kommen für die Analyse zum Einsatz:

- Lineare Regression[9]
- Random Forest[10]
- XGBoost[11]
- Multilayer Perceptron mit mehr als zwei Ebenen[12]

Die Effizienz dieser Modelle werden anhand ihres Gütemaßes dem adjusted R^2 (kurz adj. R^2) sowie der Verlustmessung Root Mean Squared Error (kurz RSME) bewertet[13]. Zudem können auf Basis der erstellten Prognosen die einflussreichsten Faktoren aller Modelle entnommen werden. Die Effizienz sowie die Faktoren der Modelle, welche einen hohen Einfluss die Prognose haben sind weitere Bestandteile der Ergebnisdiskussion neben den Prognosen.

[5] Vgl. *Rahm*, *E.*, *Do*, *H. H.*, 2000.
[6] Vgl. *Box*, *G. E. P.*, *Cox*, *D. R.*, 1964.
[7] Vgl. *Cerda*, *P.*, *Varoquaux*, *G.*, *Kégl*, *B.*, 2018.
[8] Vgl. *Bergstra*, *J.*, *Bengio*, *Y.*, 2012.
[9] Vgl. *Verbeek*, *M.*, 2017.
[10] Vgl. *Breiman*, *L.*, 2001.
[11] Vgl. *Chen*, *T.*, *Guestrin*, *C.*, 2016.
[12] Vgl. *Brause*, *R.*, 1991; *Goodfellow*, *I.*, *Bengio*, *Y.*, *Courville*, *A.*, 2017.
[13] Vgl. *Goodfellow*, *I.*, *Bengio*, *Y.*, *Courville*, *A.*, 2017; *Larose*, *D. T.*, *Larose*, *C. D.*, 2015.

Validiert werden die Ergebnisse anhand zweier Benchmarks. Diese sind zu einem der ortsübliche Mietspiegel[14], der durch die Stadt Essen erhoben worden ist, sowie den ermittelten Angebotsmieten der Zeiträume zweites Quartal 2018 bis erstes Quartal 2019, welcher durch die CBRE, einen internationalen Dienstleister aus der Immobilienwirtschaft, im LEG-Wohnungsmarktreport 2019 erhoben worden sind[15].

Nach der Analyse werden die Prognosen, die Effizienz der Modelle sowie die einflussreichsten Faktoren im fünften Teil diese Thesis in der Ergebnisdiskussion vertieft, indem diese übersichtlich aufgearbeitet und dementsprechend gegenübergestellt werden.

4 Kommentiertes Literaturverzeichnis

Amt für Stadterneuerung und Bodenmanagement - Stadt Essen (2020): Mietspiegel 2020 - für nicht preisgebundene Wohnungen in Essen, o. O.: Amt für Stadterneuerung und Bodenmanagement, 2020, S. 19

- Der Mietspiegel 2020 der Stadt Essen ist ein Leitfaden zur Ermittlung von Mietrichtwerten der Stadt Essen und deren Wohnimmobilien.

Asaftei, *Gabriel Morgan* et al. (2018): Getting ahead of the market: How big data is transforming real estate, in: Urban Land (2018), S. 6

- Diese Untersuchung, welche durch McKinsey & Company ausgeführt worden ist, behandelt die nicht-traditionellen immobilienwirtschaftliche Kennwerte und betrachtet diese auf dem amerikanischen Immobilienmarkt.

Bergstra, *James*, *Bengio*, *Yoshua* (2012): Random search for hyper-parameter optimization, in: Journal of Machine Learning Research, 13 (2012), S. 281–305, ISSN: 15324435

- Der Random Search ist ein Vorgehen, welches die optimalsten Parameter eines Modells ermittelt mit dem Ziel die besten Schätzer zu erhalten. Durch diesen Algorithmus werden vorher festgelegte Reichenweiten von Parametern durchgetestet und anhand einer Verlustmessung bewertet und auserwählt.

[14] Vgl. *Amt für Stadterneuerung und Bodenmanagement - Stadt Essen*, 2020.
[15] Vgl. *LEG*, *CBRE*, 2019.

Box, *G. E. P.*, *Cox*, *D. R.* (1964): An Analysis of Transformations, in: Journal of the Royal Statistical Society: Series B (Methodological), 26 (1964), Nr. 2, S. 211–243

- Die BoxCox-Transformation ist eine beliebte Methode für die Transformation von Daten in eine Normalverteilung. In dieser wissenschaftlichen Arbeit wird die Umsetzung dieser Methodik beschrieben.

Brause, *Rüdiger* (1991): Neuronale Netze: eine Einführung in die Neuroinformatik, o. O.: Teubner Verlag, 1991, S. 294, ISBN: 978-3-519-02247-3

- Rüdiger Brause beschreibt in diesem Buch die naturwissenschaftlichen Grundlagen von neuronalen Netzen und die daraus entstandenen Modellarchitekturen in der Informatik.

Breiman, *Leo* (2001): Random forests, in: Machine Learning, 45 (2001), Nr. 1, S. 5–32, ISSN: 08856125

- Der Random Forest, welcher ein gängiger Algorithmus aus dem Bereich des Machine Learning ist wird in dieser wissenschaftlicher Arbeit in Gänze beschrieben. Dieser Algorithmus schätzt die Prognose mit Hilfe von Entscheidungsbäumen.

Cerda, *Patricio*, *Varoquaux*, *Gaël*, *Kégl*, *Balázs* (2018): Similarity encoding for learning with dirty categorical variables, in: Machine Learning, 107 (2018), Nr. 8-10, S. 1477–1494, ISSN: 15730565, eprint: 1806.00979

- Wenn in einem Datensatz kategoriale Daten vorhanden sind, ist es für diverse Analysemethoden notwendig, dass diese in angepasste Merkmalsvektoren (eng. feature vector) in der Regel in nominale Werte umgeschrieben werden. Dies wird über Encodingmethoden ermöglicht, welche in dieser Untersuchung beschrieben werden.

Chen, *Tianqi*, *Guestrin*, *Carlos* (2016): XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Bd. 13-17-Augu, o. O., 2016, S. 785–794, ISBN: 9781450342322, eprint: 1603.02754

- Der XGBoost Algorithmus ist eine moderne Erweiterung der Entscheidungsbäume/-Ensemble für Prognosen im Bereich des Machine Learning und erfreut sich hoher Beliebtheit. Tianqi Chen und Carlos Guestrin sind die Erfinder des Algorithmus und beschreiben diesen in dieser Arbeit in voller Gänze.

Goodfellow, *Ian*, *Bengio*, *Yoshua*, *Courville*, *Aaron* (2017): Deep Learning (Adaptive Computation and Machine Learning), o. O.: The MIT Press, 2017, S. 800, ISBN: 0262035618

- Dieses Buch behandelt die Thematik rund um Deep Learning. In diesem werden die mathematischen und konzeptionellen Hintergrunde, relevante Konzepte der linearen Algebra und Wahrscheinlichkeitstheorie beschrieben, welche gebraucht werden um neuronalen Netze aufzubauen. Neben diesen Grundlagen geht dieses Buch noch weiter und beschreibt zu dem die Optimierung und Regulierung der diversen Arten von neuronale Netzen.

Larose, *Daniel T.*, *Larose*, *Chantal D.* (2015): Data mining and predictive analytics, Second Edition, o. O.: Wiley, 2015, S. 794, ISBN: 978-1-118-11619-7

- Data Mining findet täglich mehr Verwendung und Akzeptanz, denn es ermöglicht Unternehmen profitable Muster und Trends aus ihren bestehenden Datenbanken aufzudecken und diese zu nutzen. Daniel T. und Chantal D. Larose beschreiben verschiedene Analysemethoden für ein effizientes Data Mining.

LEG, *CBRE* (2019): LEG-Wohnungsmarktreport 2019, o. O.: LEG Immobilien AG, 2019, S. 33

- Die LEG Immobilien AG bringt in regelmäßigen Abständen in Zusammenarbeit mit der CBRE den LEG-Wohnungsmarktreport heraus. Dieser Report analysiert die aktuelle Situation der Wohnimmobilienmärkte bezogen auf die Großstädte der Bundesrepublik Deutschland.

Mayring, *Philipp* (2001): Combination and Integration of Qualitative and Quantitative Analysis, in: Forum: Qualitative Social Research, 2 (2001), Nr. 1, ISSN: 1438-5627

- In dieser Arbeit stellt Phillip Mayring verschiedene Hybridmodelle für empirische Untersuchungen vor. Das Vertiefungsmodell, welcher der Rahmen dieser Thesis ist wird darin beschrieben.

Rahm, *Erhard*, *Do*, *Hong Hai* (2000): Data cleaning: Problems and current approaches, in: IEEE Data Eng. Bull. 23 (2000), Nr. 4, S. 3–13, ISSN: 03064379

- Die Datenbereinigung gehört zu den zeitintensivsten Schritten in der Modellentwicklung. In dieser Arbeit werden Methoden und Ansätze zur Datenbereinigung für wiederkehrende Probleme beschrieben.

Stadt Essen, *InWIS Forschung* (2018): Wohnungsnachfrageanalyse Essen 2025+, o. O.: InWIS Forschung & Beratung, 2018, S. 84

- Dieses Gutachten, welches durch die Stadt Essen in Auftrag gegeben worden ist, behandelt die Frage nach dem Wohnungsangebot sowie -nachfrage der Stadt Essen in den kommenden Jahren und skizziert mögliche Handlungsfelder um den steigenden Bedarf zu minimieren.

Verbeek, *Marno* (2017): A guide to modern econometrics, o. O.: John Wiley & Sons, 2017, S. 508, ISBN: 978-1-119-40115-5

- Dieses Buch dient als Leitfaden für alternative Techniken in der Ökonometrie mit Schwerpunkt auf Intuition und der praktischen Umsetzung dieser Ansätze. Es deckt ein breiten Themenspektrum ab einschliesslich der linearen Regression mit dessen Diagnostik, Zeitreihenanalysen und Paneldatenanalysen.

Wirth, *Rüdiger* (2000): CRISP-DM : Towards a Standard Process Model for Data Mining, in: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (2000), Nr. 24959, S. 29–39

- Diese wissenschaftliche Arbeit beschreibt den kompletten CRISP-DM Prozess mit allen zugehörigen Prozessschritten und deren Inhalten.

5 Vorläufige Gliederung

Nr.	Kapitel	Behandelt
1	**Einleitung** • 1.1 Motivation • 1.2 Problemstellung • 1.3 Zielsetzung	• Ziel- und Problembeschreibung des Themas
2	**Der Immobilienmarkt in Essen, NRW** • 2.1 Aktuelle und prognostizierte kommunale Entwicklungen der Mietsituation und des Mietbedarfes der Stadt Essen • 2.2 Beschreibung der Unterschiede der traditionellen und nicht-traditionellen Faktoren, welche die Grundlage der Analyse bilden	• Grundlagenbeschreibung • Spiegelt den Bereich „Business Understanding“ der CRISP-DM Methology ab
3	**Der Datensatz und dessen Aufbereitung** • 3.1 Beschreibung der Datensätze sowie der Benchmarks, welche zur Analyse genutzt werden • 3.2 Beschreibung der Datenbereinigung und Zusammenschluss der traditionellen und nicht-traditionellen Faktoren • 3.3 Explorative Datenanalyse • 3.4 Präparation der Daten für die Anwendungen zur Prognose	• Beschreibung und Aufarbeitung aller genutzten Daten, welche die Basis für die Prognosen bilden • - Spiegelt die Bereiche „Data Understanding“ sowie „Data Preparation“ der CRISP-DM Methology ab

Fortführung auf der nächsten Seite

Nr.	Kapitel	Behandelt
4	**Modellierung und Analyse** • 4.1 Beschreibung der verwendeten Modelle sowie deren Einsatz zur Prognose der Mieten – Lineare Regression – Random Forest – XGBoost – Multilayer Perceptron • 4.2 Vergleich der Güte sowie der Verlustmessungen	• Weiterverarbeitung aller genutzten Daten, sowie eine erste Analyse, durch den Einsatz der Prognosemodelle • Spiegelt den Bereich „Modeling“ der CRISP-DM Methology ab
5	**Evaluierung und Diskussion** • 5.1 Aufbereitung der Validierungsergebnisse mit den verschiedenen Prognosen sowie Gegenüberstellung mit dem Benchmark • 5.2 Betrachtung der effizientesten Faktoren, welche durch die Prognosemodelle für die Validierung genutzt worden sind • 5.3 Diskussion über die Ergebnisse	• Behandelt die Validierung und diskutiert diese mit Betrachtung auf die genutzten Prognosemodelle und vergleicht diese mit den Benchmarks, sowie deren Faktoren • Spiegelt den Bereich „Evaluation“ der CRISP-DM Methology ab
6	**Fazit**	• Schlussfolgerung aus der Evaluierung und Diskussion

Literaturverzeichnis

Amt für Stadterneuerung und Bodenmanagement - Stadt Essen (2020): Mietspiegel 2020 - für nicht preisgebundene Wohnungen in Essen, o. O.: Amt für Stadterneuerung und Bodenmanagement, 2020, S. 19

Asaftei, *Gabriel Morgan*, *Doshi*, *Sudeep*, *Means*, *John*, *Sanghvi*, *Aditya* (2018): Getting ahead of the market: How big data is transforming real estate, in: Urban Land (2018), S. 6

Bergstra, *James*, *Bengio*, *Yoshua* (2012): Random search for hyper-parameter optimization, in: Journal of Machine Learning Research, 13 (2012), S. 281–305

Box, *G. E. P.*, *Cox*, *D. R.* (1964): An Analysis of Transformations, in: Journal of the Royal Statistical Society: Series B (Methodological), 26 (1964), Nr. 2, S. 211–243

Brause, *Rüdiger* (1991): Neuronale Netze: eine Einführung in die Neuroinformatik, o. O.: Teubner Verlag, 1991, S. 294

Breiman, *Leo* (2001): Random forests, in: Machine Learning, 45 (2001), Nr. 1, S. 5–32

Cerda, *Patricio*, *Varoquaux*, *Gaël*, *Kégl*, *Balázs* (2018): Similarity encoding for learning with dirty categorical variables, in: Machine Learning, 107 (2018), Nr. 8-10, S. 1477–1494

Chen, *Tianqi*, *Guestrin*, *Carlos* (2016): XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Bd. 13-17-Augu, o. O., 2016, S. 785–794

Goodfellow, *Ian*, *Bengio*, *Yoshua*, *Courville*, *Aaron* (2017): Deep Learning (Adaptive Computation and Machine Learning), o. O.: The MIT Press, 2017, S. 800

Larose, *Daniel T.*, *Larose*, *Chantal D.* (2015): Data mining and predictive analytics, Second Edition, o. O.: Wiley, 2015, S. 794

LEG, *CBRE* (2019): LEG-Wohnungsmarktreport 2019, o. O.: LEG Immobilien AG, 2019, S. 33

Mayring, *Philipp* (2001): Combination and Integration of Qualitative and Quantitative Analysis, in: Forum: Qualitative Social Research, 2 (2001), Nr. 1

Rahm, *Erhard*, *Do*, *Hong Hai* (2000): Data cleaning: Problems and current approaches, in: IEEE Data Eng. Bull. 23 (2000), Nr. 4, S. 3–13

Stadt Essen, *InWIS Forschung* (2018): Wohnungsnachfrageanalyse Essen 2025+, o. O.: InWIS Forschung & Beratung, 2018, S. 84

Verbeek, *Marno* (2017): A guide to modern econometrics, o. O.: John Wiley & Sons, 2017, S. 508

Wirth, *Rüdiger* (2000): CRISP-DM : Towards a Standard Process Model for Data Mining, in: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (2000), Nr. 24959, S. 29–39