Digitale Fußabdrücke im Kredit-Scoring – Ein Vergleich von traditionellen statistischen und Machine Learning Verfahren

Tobias Berg | Maximilian Kreft

Eine von FIRM geförderte Forschung [vgl. Berg et al. 2019] zeigt, dass „digital Footprint"-Variablen maßgeblich mit der Ausfallwahrscheinlichkeit von Kunden zusammenhängen und somit eine wertvolle alternative Datenquelle zur Risikoklassifizierung im Privatkundengeschäft darstellen. Neben dem Informationswert der Kreditnehmermerkmale kann jedoch auch das Klassifizierungsverfahren zur Modellierung des empirischen Zusammenhanges einen entscheidenden Einfluss auf die Trennschärfe eines Scoring-Modells haben.

Während die Bankpraxis vor allem aufgrund der Einfachheit, Erklär- und Validierbarkeit das logistische Regressionsmodell als Marktstandard für Kredit-Scoring bevorzugt, zeigen aktuelle akademische Vergleichsstudien, dass Klassifizierungsverfahren aus dem Bereich des Maschinellen Lernens („machine learning") häufig eine höhere Voraussagekraft erzielen [vgl. Lessmann et al. 2015]. Aufbauend auf den Ergebnissen von Berg et al. [vgl. Berg et al. 2019], haben wir daher untersucht, ob sich die Trennschärfe einer „Digital Footprint"-Scorekarte mittels Klassifizierungsverfahren aus dem Bereich des Maschinellen Lernens im Vergleich zur logistischen Regression verbessern lässt [vgl. Kreft 2019].

Die Studie beschäftigt sich insbesondere mit Klassifizierungsverfahren aus dem Bereich des Maschinellen Lernens, die Entscheidungsbäume als Grundlage zur Vorhersage von Kreditausfällen heranziehen. Entscheidungsbäume können nicht-lineare Zusammenhänge abbilden. Sie sind aufgrund ihrer hohen Flexibilität allerdings anfällig für eine Überanpassung an den Entwicklungsdatensatz („overfitting"). Klassifizierungsverfahren wie Random Forest [vgl. Breiman 2001] und XGBoost [vgl. Chen und Guestrin 2016] sind beliebte Algorithmen, die robuste Ausfallwahrscheinlichkeiten mittels einer Vielzahl von möglichst unkorrelierten Entscheidungsbäumen schätzen. Sie minimieren auf diese Weise das Risiko einer Überanpassung an die Daten (siehe ▶ Abb. 01).

Der für die Studie verwendete Datensatz besteht aus 10 Variablen des „Digital Footprint", wie beispielsweise dem E-Mail-Anbieter, dem Betriebssystem oder der Schreibweise beim Ausfüllen von Online-Formularen [siehe Berg et al. 2019 für eine weitergehende Beschreibung]. Basierend auf der Analyse dieses umfangreichen Datensatzes von über 250.000 Bestellungen eines deutschen E-Commerce-Unternehmens stellen wir zunächst fest, dass Ran-

Abb. 01: Random Forest versus XGBoost

Random Forest	XGBoost
Parallele homogene Ensemblemethode basierend auf der Idee des „Bagging" I. **Bagging** („Bootstrapped Aggregation") entwickelt ein Ensemble von Entscheidungsbäumen basierend auf zufälligen Stichproben mit Zurücklegen („bootstrapped samples") II. **Random Forest** ist eine Weiterentwicklung des Bagging Algorithmus, welche mittels „Random Feature Selection" die Entscheidungsbäume weiter de-korreliert.	Sequenzielle homogene Ensemblemethode basierend auf der Idee des „Stochastic Boosting" I. **Gradient Boosting** entwickelt ein Ensemble von Entscheidungsbäumen indem jeder zusätzliche Entscheidungsbaum zum Restfehler des vorherigen Entscheidungsbaumes angepasst wird II. **XGBoost** ist eine Weiterentwicklung des Gradient Boosting Algorithmus, welche nicht nur auf zufälligen Stichproben ohne Zurücklegen („random subsampling") basiert sondern auch mittels „Random Feature Selection" die Entscheidungsbäume weiter de-korreliert.

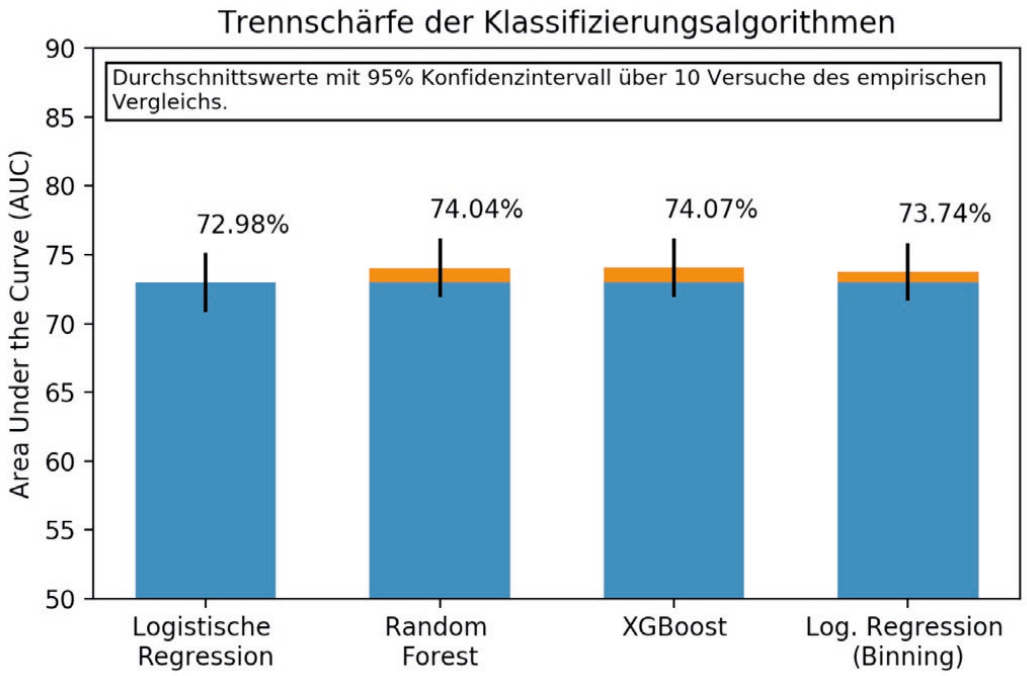
Quelle: Eigene Darstellung basierend auf Breiman (1996 und 2001), Friedman (2002) und Chen und Guestrin (2016)

dom Forest und XGBoost die Trennschärfe der „Digital Footprint"-Scorekarte in moderatem Umfang verbessern (siehe ▶ Abb. 02). Quantitativ verbessert sich die Trennschärfe – gemessen anhand der Messgröße „Area Under the Curve" (AUC) – bei beiden Methoden (Random Forest, XGBoost) um ca. einen Prozentpunkt gegenüber der logistischen Regression. Diese empirischen Ergebnisse stehen in Einklang mit der führenden akademischen Literatur im Bereich Kredit-Scoring [vgl. Lessmann et al. 2015, Xia et al. 2017] und bekräftigen zugleich, dass Random Forest und XGBoost insbesondere auch für Kreditportfolien mit niedrigen Ausfallraten geeignet sind [vgl. Brown und Mues 2012]. Diese moderate Steigerung der Trennschärfe durch die Klassifizierungsverfahren könnte in der Risikoklassifizierung im Privatkundengeschäft durchaus ökonomische Relevanz haben. Im Vergleich zeigt sich jedoch die höhere Bedeutung des Informationswert der verwendeten Kreditnehmermerkmale: Die Trennschärfe eines Kredit-Scores einer Kreditauskunftei lässt sich im untersuchten Anwendungsfall nur durch Hinzufügen der Variable „E-Mail-Anbieter" des „Digital Footprint" bereits um mehr als zwei Prozentpunkte anheben [vgl. Berg et al. 2019].

Des Weiteren können wir auch zeigen, dass die Trennschärfe der logistischen Regression erheblich verbessert werden kann, wenn die strikte Annahme der Linearität im Modellentwicklungsprozess angemessen berücksichtigt wird. Aufgrund der geringen Anzahl an Kreditnehmermerkmalen im zugrundeliegenden Datensatz ist eine entsprechende Analyse der Datenstruktur zur Überprüfung der Einhaltung der Modellannahmen mit geringem Aufwand möglich. Bereits die Anwendung einfachster Datenvorverarbeitungsschritte („data pre-processing") wie die „Binning" Methode, d.h. die Übersetzung kontinuierlicher in ordinale Kreditmerkmale, führen zu signifikanten Verbesserungen der Trennschärfe um 0,8 Prozentpunkte (siehe ▶ Abb. 02). Solche Datenvorverarbeitungsschritte sind in der Praxis weitverbreitet. Im Ergebnis können wir ein konkurrenzfähiges „Digital Footprint"-Scorekarte mittels logistischer Regression erhalten, welches sich statistisch nicht signifikant von den Modellen basierend auf Random Forest und XGBoost unterscheidet. Diese Erkenntnis unterstützt die These, dass die größten Fortschritte in der Voraussagekraft – basierend auf dem Gesetz des abnehmenden Ertragszuwachses – von einfacheren Modellansätzen erzielt werden [vgl. Hand 2006]. Es ist zweifelhaft, ob der geringfügige Fortschritt in der Voraussagekraft durch die komplexeren Modellansätze des Maschinellen Lernens in unserer empirischen Studie die zusätzlichen Modellrisiken hinsichtlich der Interpretier- und Validierbarkeit der Kredit-Scorekarte rechtfertigen.

Letztendlich können die empirischen Ergebnisse dieser Studie aber Aufschluss über die Bedeutung der Datenstruktur für die Wahl eines geeigneten Klassifizierungsverfahren geben. Lineare statistische Modelle erscheinen demnach gut geeignet Kredit-Scoring-Modelle bei Anwendungsfällen mit strukturierten Datensätzen zu entwickeln, welche wenige Kreditnehmermerkmale („low dimensional datasets") aufweisen. In diesen Fällen kann die weitestgehende Einhaltung der strikten Annahmen linearer Modelle hinsichtlich nicht-linearer Abhängigkeiten und Wechselwirkungen zwischen Kreditnehmermerkmalen mit vergleichsweise geringem Aufwand durch Datenvorverarbeitungsschritte sichergestellt werden. Maschinelle Lernverfahren können dahingehend als Benchmark unterstützend in der Modellentwicklungspraxis genutzt werden, um den potenziellen Nutzen einer eingehenden Datenvoranalyse zu quantifizieren.

Abb. 02: Trennschärfe der Klassifizierungsalgorithmen



Quelle: Eigene Darstellung basierend auf Kreft (2019)

Diese Schlussfolgerung sollte allerdings nicht als allgemeine Aussage zum praktischen Nutzen von Klassifizierungsverfahren aus dem Bereich des Maschinellen Lernens verstanden werden. Jedoch erscheint eine Anwendung dieser Modellansätze im Bereich Kredit-Scoring nur in Verbindung mit höher dimensionalen und komplexeren Datenstrukturen („Big Data") gerechtfertigt. In diesen Fällen ist die notwendige Analyse nicht-linearer Zusammenhänge und Wechselwirkungen zwischen den Merkmalen nur unter äußerst hohem Aufwand möglich. Hier können Modellansätze aus dem Bereich des Maschinellen Lernens Abhilfe schaffen. Jedoch stellt die Komplexität dieser Modellansätze neue Herausforderungen an die Interpretier- und Validierbarkeit. Daher sollten Ansätze zur Erklärbarkeit im Mittelpunkt der Forschung stehen, um Probleme bezogen auf rechtliche und operative Hindernisse und Modellrisiken zu adressieren („Explainable AI").

Fazit

Wir haben festgestellt, dass der Mehrwert von Klassifizierungsverfahren aus dem Bereich des Maschinellen Lernens im traditionellen Kredit-Scoring mit wenigen Kreditnehmermerkmalen fragwürdig erscheint. Zwar verbessern die Modellansätze, wie Random Forest und XGBoost, die Trennschärfe der Vorhersage der Zahlungsausfälle von Verbrauchern gegenüber traditionellen statistischen Verfahren. Jedoch lässt sich eine vergleichbare Verbesserung der Güte des Kredit-Scoring Modells basierend auf traditionellen statistischen Verfahren schon durch die explizite Modellierung von nicht-linearen Abhängigkeiten und Wechselwirkungen zwischen den Merkmalen erzielen. Soweit die Einfachheit der Datenstruktur eine Analyse dieser Abhängigkeiten und Wechselwirkungen zulässt, erscheint die Notwendigkeit für den Einsatz Maschinellen Lernens im Bereich Risikoklassifizierung nicht zwingend gegeben.

Der potenzielle Einsatz maschinellen Lernens geht daher insbesondere mit der Erschließung alternative Datenquellen aus dem Bereich Big Data einher. In dieser Hinsicht sollte ein besonderes Augenmerk der zukünftigen Forschung auf der Erklärbarkeit und Validierbarkeit dieser Klassifizierungsverfahren liegen.

Literatur

Berg, T./Burg, V./Gombović, A./Puri, M. [2019]: On the Rise of FinTechs: Credit Scoring using Digital Footprints, in: Review of Financial Studies (forthcoming)

Breiman, L. [1996]: Bagging Predictors, in: Machine Learning 24/2, S. 123–140

Breiman, L. [2001]: Random Forests, in: Machine Learning 45/1, S. 5-32

Brown, I. und Mues, C. [2012]: An experimental comparison of classification algorithms for imbalanced credit scoring data sets, in: Expert Systems with Applications 39/3, S. 3446–3453.

Chen, T. und Guestrin, C. [2016]: XGBoost: A Scalable Tree Boosting System, in: KDD2016: 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, S. 785-794

Friedman, J. H. [2002]: Stochastic gradient boosting, in: Computational Statistics and Data Analysis 38/4, S. 367–378

Hand, D. J. [2006]: Classifier Technology and the Illusion of Progress, in: Statistical Science 21/1, S. 1–14

Kreft, M. [2019]: Credit Scoring with Alternative Data: A Comparison of Traditional Statistical and Machine Learning Methods, Masterarbeit (Frankfurt School of Finance and Management)

Lessmann, S. / Baesens, B. / Seow, H.-V. and Thomas, L. C. [2015]: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, in: European Journal of Operational Research 247/1, S. 124–136

Xia, Y. / Liu, C. / Li, Y. and Liu, N. [2017]: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, in: Expert Systems with Applications 78, S. 225–241.

Autoren

Prof. Dr. Tobias Berg

Associate Professor of Finance,
Frankfurt School of Finance & Management,
Frankfurt am Main

Maximilian Kreft

Master of Finance Graduate,
Frankfurt School of Finance & Management,
Frankfurt am Main