

# Brezel-Cast: Verkaufsprognose von Backwaren

Nico Döring<sup>1</sup>, Jonathan Kreiss<sup>1</sup>, Thomas Schuster <sup>1</sup> und Raphael Volz <sup>1</sup>

**Abstract:** In diesem Papier diskutieren wir die Anwendbarkeit von Verfahren der künstlichen Intelligenz zur Prognose von Absatzzahlen für eine Bäckerei mittlerer Größe. Dabei wird beschrieben, wie bei der Entwicklung zusätzliche Daten (Kontextinformationen) zur Prognose genutzt werden. Daraufhin werden zwei Verfahren des maschinellen Lernens trainiert und im Ergebnis miteinander verglichen. Neben einer abschließenden Bewertung und Ausblick auf zukünftige Verbesserungen, wird zudem eine Einschätzung zum Einsatz im Produktivbetrieb abgegeben.

**Keywords:** Künstliche Intelligenz, Maschinelles Lernen (ML), Absatzprognose, Nachhaltige Produktion

## 1 Einleitung

Künstliche Intelligenz (KI) durchdringt immer mehr Bereiche unseres Lebens und Unternehmen setzen zunehmend auf KI-Verfahren wie das maschinelle Lernen (ML). Das Potential von ML-Verfahren ist neben der Verbesserung von Produkten und Dienstleistungen auch eine Steigerung der Prozesseffizienz. Während viele Anwendungsfälle derzeit von Großunternehmen beschrieben werden, setzen inzwischen auch zunehmend mittelständische Betriebe auf ML. Ein typischer Anwendungsfall für den Einsatz von ML ist die Prognose künftigen Kundenverhaltens wie zum Beispiel von Verkaufszahlen bestimmter Produkte. Besonders bei schnelllebigem Konsumgütern wie Nahrungsmitteln, bringt dies viele Vorteile mit Blick auf eine effiziente, nachhaltige Produktion und Logistik. Dies gilt nicht nur für große Supermarktketten, sondern auch für kleinere Betriebe (Bäckereien). In diesem Papier zeigen wir, wie Verfahren des maschinellen Lernens (ML) in einer Bäckerei eingesetzt werden kann, um Absatzzahlen prognostizieren zu können. Wir zeigen, wie die Verkaufsmenge einzelner Produkte pro Filiale prognostiziert und so die Menge von Retouren und Abverkäufe reduziert werden kann, um die Produktion nachhaltiger zu gestalten und Ausschuss zu reduzieren.

---

<sup>1</sup> Hochschule Pforzheim, Tiefenbronner Straße 65, 75175 Pforzheim,  
{doeringi, kreissjo, thomas.schuster, raphael.volz}@hs-pforzheim.de,

 <https://orcid.org/0000-0002-9539-1627>, <https://orcid.org/0000-0003-2816-1618>

## 2 Verwandte Arbeiten

Die Vorhersage von Lebensmitteln und auch von Backwarenverkäufen ist eine Aufgabe der Zeitreihenprognose. Im Vergleich zu traditionellen Methoden zur Verkaufsprognose, können Ansätze des maschinellen Lernens eine höhere Prognosegenauigkeit, Flexibilität und Skalierbarkeit über große Datensätze bieten [Ta19]. [Ne10] vergleichen empirisch ML-Ansätze für Zeitreihenprognosen. Im Vergleich performen Neuronale Netze und Gaussian Process Regression am besten. K-Nearest Neighbour erzielt hingegen eine mittelmäßige Qualität der Prognose, zeigt im Vergleich aber einen geringeren Rechenaufwand. Aufgrund der Aufwände und der Qualität bei der Verkaufsprognose von salzigen Snacks eines Supermarktes in [Ba15] kann sich Random Forest als Verfahren im Vergleich durchsetzen. [BS11] zeigen erfolgreich die Anwendbarkeit von Entscheidungsbaumalgorithmen zur Prognose der Nachfrage in Food Courts. In [LI17] werden basierend auf Deep Learning und meteorologischen Daten Verkäufe von Lebensmitteln einer japanischen Supermarktkette prognostiziert. [SV19] nutzen ebenfalls Kontextinformationen (z.B. Wetterdaten, Ferientage) um die Nachfrage von Parkplätzen in Parkhäusern zu prognostizieren.

## 3 Ausgangssituation und Prognoseansatz

Gemeinsam mit einer mittelständischen Bäckerei, die derzeit mehrere Filialen betreibt und durch den zentralen Produktionsstandort täglich frisch beliefert, sollte untersucht werden wie genau Absatzzahlen für einzelne Produkte pro Filiale prognostiziert werden können und ob diese Vorhersage besser ist als die der eigenen Mitarbeiter. Somit kann die Produktion der gesamten Kette effizient auf die Nachfrage ausgerichtet werden – also sowohl Über- als auch Unterproduktion reduziert werden. Im Projekt wurden exemplarisch die Verkäufe von Brezeln analysiert. Die zentrale Forschungsfrage war daher: Kann ein ML-Modell die Verkaufsmenge von Brezeln für den nächsten Tag, pro Filiale genauer prognostizieren als der Bäckermeister?

Um diese Frage zu beantworten, wurden Algorithmen aus dem überwachten maschinellen Lernen evaluiert. Die Datengrundlage waren Verkaufszahlen der letzten vier Jahre (2016-2019), welche aus dem Kassensystem einer Filiale der Bäckerei extrahiert wurden. Im Hinblick auf die Algorithmen und durch Literaturanalyse ähnlicher Fragestellungen, wurden Regressionsansätze, Entscheidungsbaumalgorithmen (besonders Random Forest Verfahren) und Neuronale Netze als vielversprechend erachtet. Für die Prognose wurde untersucht wie geeignet die beiden Algorithmen K-Nearest Neighbour (kNN) und Random Forest sind, um die Verkaufsmenge der Brezeln vorherzusagen.

## 4 Konzeption und Entwicklung der Absatzprognose

In diesem Abschnitt beschreiben wir die Entwicklung der Absatzprognose über die

komplette Verarbeitungstrecke der Daten. Dazu gehen wir zuerst auf die Vorverarbeitung der Daten ein und vergleichen anschließend die beiden Ansätze miteinander. Der Abschnitt endet mit einer abschließenden Bewertung der Modelle.

#### **4.1 Datenaufbereitung und Modellierung**

Zunächst wurden die Daten des Kassensystems (CSV-Dateien) mit Hilfe von pandas eingelesen. Der Datensatz umfasste Verkaufszahlen einer Filiale im Zeitraum von 2016 bis 2019. Die Parameter des Datensatzes wurden gefiltert und auf, für die Prognose bedeutenden, reduziert. Dies waren: Verkaufsdatum und zugehörige Verkaufsmenge jeder Transaktion. Anschließend wurden Ausreißer und Stornierungen (negative Verkaufswerte) bereinigt. Zusätzlich wurden Transaktionen eines Tages aggregiert und zu  $n = 1461$  ( $365 * 4 + 1$ ) Beobachtungen aufsummiert.

Um die Verkaufszahlen besser prognostizieren zu können, wurde der Datensatz um Kontextinformationen angereichert, also durch weitere, erklärende Variablen erweitert. Dazu gehörte beispielsweise Informationen über den jeweiligen Wochentag, ob es sich an einem Tag um einen Ferientag handelt und Wetterdaten. Die Wetterdaten (durchschnittliche Tagestemperatur und Niederschlagsmenge pro Tag) wurden über das Informationsangebot des Deutschen Wetterdienstes (DWD) [DWD21] bezogen, welches historische und auch regionale Wettermessungen kostenlos bereitstellt.

#### **4.2 Datenanalyse**

Zur Bewertung der unabhängigen Variablen wurden zunächst einige Visualisierungen vorgenommen. Zunächst wurde die Verkaufsmenge pro Tag anhand einiger Diagramme wie z.B. Histogramm, Liniendiagramm und Boxplots visualisiert. Das Ziel war es die Verkaufsdaten zu untersuchen und Trends, wie zum Beispiel Saisonalitäten zu entdecken. Anhand eines Histogramms wurde beispielsweise festgestellt, dass an den meisten Tagen, zwischen 60 und 175 Brezeln verkauft werden.



Abb. 1: Spearman Korrelationskoeffizient Verkaufsmenge zu Parametern

Anschließend wurde mit Hilfe des Spearman Koeffizienten die Korrelation zwischen der Verkaufsmenge und den einzelnen Parametern ermittelt (siehe Abb. 1). Erwartungsgemäß zeigte sich die höchste Korrelation (0,5) zwischen dem Wochentag und der Verkaufsmenge. Diese entspricht einer positiven mittelstarken Korrelation. Der Boxplot in Abbildung 2 verdeutlicht ebenfalls diesen Zusammenhang. Wie erwartet ist die Verkaufsmenge an Freitagen (4), Samstagen (5) und Sonntagen (6) höher als an anderen Wochentagen. Ein ähnliches Bild zeigte sich an Feiertagen.

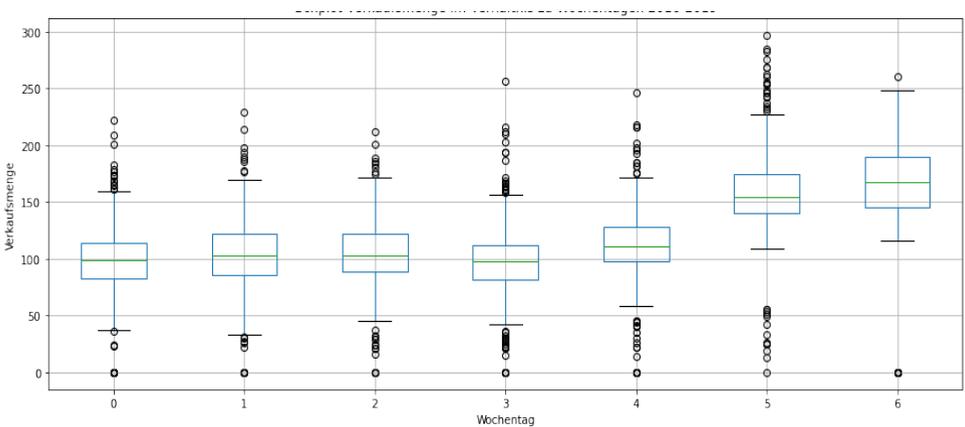


Abb. 2: Brezel Verkaufsmenge pro Wochentag von Montag (0) - Sonntag (6)

### 4.3 Modell Training

Das Trainingsverfahren wurde, durch ein Sliding Window erweitert [HHS17]. Dabei

wird ein Zeitfenster rollierend über dem Datensatz verschoben und durch die aus dem Fenster entnommenen Datenpunkte (Lag Variablen) erweitert. Dieser Ausschnitt an Vergangenheitswerten liefert somit zusätzliche Information für eine Prognose.

**K-Nearest-Neighbour:** Die Daten wurden in Trainings- und Testdaten aufgeteilt und durch eine Datenskalierung transformiert. Da der kNN Algorithmus nur den Hyperparameter  $k$  zum Tuning bereitstellt, wurde mit Hilfe des GridSearchCV-Algorithmus der optimale Wert  $k=4$  ermittelt. Nach Training und Prognose der Werte aus dem Testdatensatz ergaben sich die Evaluierungsparameter aus Tabelle 1.

Evaluierungsparameter	Wert
Mean Absolute Error (MAE)	21.07
Prognosegenauigkeit in % aus MAE	78.64 %
Mean Squared Error (MSE)	768.89
Root Mean Squared Error (RMSE)	27.73

Tab. 1: Gemittelte Out-of-Sample Fehlerrate des kNN Modells

**Random Forest:** Die Daten wurden auch hier aufgeteilt. Um die Hyperparameterauswahl zu optimieren und eine Kreuzvalidierung zu ermöglichen, wurde eine randomisierte Suche (RandomSearchCV) und in einem zweiten Schritt der GridSearchCV-Algorithmus auf die eingegrenzten Parametermöglichkeiten angewandt. Als Baseline diente eine Random Forest Prognose mit den Standard-Hyperparametern (Default Modell) aus der scikit-learn Bibliothek. Die Prognose mit dem Default Modell erreichte eine Prognosegenauigkeit von 83.85%. Mit den Parametern aus der randomisierten Suche verbesserte sich das Ergebnis um 1.27% auf 85.12%. Eine Auswahl durch den GridSearchCV-Algorithmus, führte zu keiner weiteren Steigerung. Daher wurde die Annahme gestellt, dass das Nutzenoptimum erreicht wurde. Die wichtigsten Hyperparameter des Modells am Nutzenoptimum (Finales Modell) sind 1400 klassifizierende Bäume und eine maximale Baumtiefe von 10. Die Ergebnisse sind in Tabelle 2 zusammengefasst.

Evaluierungsparameter	Default Modell	Finales Modell	Verbesserung
Mean Absolute Error (MAE)	18.71	16.98	1.73
Prognosegenauigkeit in % aus MAE	83.85%	85.12%	1.27%
Mean Squared Error (MSE)	603.34	495.17	108.17
Root Mean Squared Error (RMSE)	24.56	22.25	2.31

Tab. 2: Gemittelte Out-of-Sample Fehlerrate des Random Forest Modells

#### 4.4 Evaluation

Aus unserem Vergleich der Machine Learning Algorithmen hat das Finale Modell des Random Forest mit einer Prognosegenauigkeit von 85.12% am besten abgeschnitten. Der kNN-Algorithmus konnte nur eine Prognosegenauigkeit von 78.64% liefern. Im Vergleich ist somit die Prognose des Random Forrest im Schnitt um 6.48% besser. Auch konnte das MSE beim Random Forest deutlich gesenkt werden, was für die Robustheit

des Modells spricht. Die Ergebnisse können nun genutzt werden, um weitere Anpassungen vorzunehmen und die Prognosegenauigkeit des Random Forests zu verbessern. Zudem kann es als Baseline für Vergleiche mit weiteren Verfahren dienen.

Evaluierungsparameter	kNN	Random Forest	Verbesserung
Mean Absolute Error (MAE)	21.07	16.98	4.09
Prognosegenauigkeit in % aus MAE	78.64%	85.12%	6.48%
Mean Squared Error (MSE)	768.89	495.17	273.72
Root Mean Squared Error (RMSE)	27.73	22.25	5.48

Tab. 3: Vergleich kNN zu Random Forest

## 5 Fazit und Ausblick

Zur Prognose des Absatzes von Backwaren wurden in diesem Projekt zwei Algorithmen, der Random-Forest-Algorithmus und der K-Nearest-Neighbour, untersucht, getestet und miteinander verglichen. Das Ziel war es die Verkaufsmenge der Brezeln, einer Bäckereifiliale möglichst genau für einen Tag vorherzusagen. In ersten Analysen kristallisierte sich heraus, dass der Wochentag einen starken Einfluss auf die Verkaufsmenge hat. Nachdem die Daten entsprechend aufbereitet und mit Kontextinformationen (Wochen- und Feiertage sowie Wetterdaten) angereichert wurden, wurden die Modelle trainiert. Nach Hyperparameteroptimierung und Evaluation der beiden getesteten Algorithmen, erzielte der Random Forest-Algorithmus das beste Ergebnis, mit einer Prognosegenauigkeit von über 85% und einer Abweichung von nur 17 Brezeln gegenüber den tatsächlichen Absatzzahlen. Von der Bäckerei wissen wir, dass man sich bei einer manuellen Prognose etwa um die doppelte Menge verschätze. Somit könnte künftig die manuelle Einschätzung der Mitarbeiter ersetzt werden. In einem nächsten Schritt können weitere Verfahren trainiert und mit den bereits trainierten Modellen verglichen werden. Darüber hinaus lässt sich die Verkaufsprognose nun auch auf andere Artikel und Filialen sowie zusätzliche Verkaufsstrategien erweitern. Die interessanteste Erweiterung wäre die direkte Anbindung an das Kassensystem, sodass die Prognose mit dem Sliding Window Ansatz jeweils tagesaktuell mit den entsprechenden Verkaufsdaten durchgeführt werden kann. Für Prognosen in der Praxis ist die Komplexität des eingesetzten Verfahrens, Modells und Benutzerschnittstelle abzuwägen. So kann ein leicht erklärbares Verfahren die Akzeptanz und das Vertrauen in das System fördern. Eine Reduzierung von Features vereinfacht zudem die Verwaltung von Datenabhängigkeiten. Die Benutzerschnittstelle soll einfach gehalten werden, z.B. jeweils eine Spalte mit dem historischen Durchschnitt der Verkaufszahlen des Produkts pro Wochentag und der angepassten KI-Prognose.

## Literaturverzeichnis

- [Ba15] Bajari, P.; Nekipelov, D.; Ryan, S. P.; Yang, M.: Machine learning methods for demand estimation. *American Economic Review*, 105 (5), 481-85, 2015.

- 
- [BS11] Bozkir, A. S.; Sezer, E. A.: Predicting food demand in food courts by decision tree approaches. *Procedia Computer Science* Vol. 3, pp. 759-763, 2011.
- [DWD21] DWD, Deutscher Wetterdienst, <https://opendata.dwd.de/>, zugegriffen: 17/05/2021.
- [HHS17] Hota, H. S.; Handa, R.; Shrivastava, A. K.: Time Series Data Prediction Using Sliding Window Based RBF Neural Network. In: *International Journal of Computational Intelligence Research* Vol. 13 No. 5, pp. 1145-1156, 2017.
- [LI17] Liu, X. und Ichise, R.: Food Sales Prediction with Meteorological Data – A Case Study of a Japanese Chain Supermarket. In: *Data Mining and Big Data: Second International Conference. Lecture Notes in Computer Science* Vol. 10387, Springer International Publishing AG, Fukuoka Japan, pp. 93-104, 2017.
- [Ne10] Nesreen, K. A. et al.: An Empirical Comparison of Machine Learning Models for Time Series Forecasting, *Econometric Reviews* 29:5-6, pp. 594-621, 2010.
- [SV19] Schuster, T. und Volz, R.: Predicting Parking Demand with Open Data. In: *Digital Transformation for a Sustainable Society in the 21st Century. I3E 2019. Lecture Notes in Computer Science* Vol. 11701, Springer, Cham, pp. 218-229, 2019.
- [Ta19] Tarallo, E. et al.: Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. In: *IFAC-PapersOnLine* Vol. 52 Issue 13, pp. 737-742, 2019.