Isolation Forest – Anwendung von Anomalie-Erkennung im Kalibrierwesen

333. PTB-SEMINAR 06.05.2026 – PHILIPP ALFTER

04.05.2026

Motivation

Anwendung von KI in der Industrie

- Zunehmend in der Fertigung
  - Automatisierte Fehlererkennung
  - KI-gestützte Werkstoffprüfung

Motivation

KI im Kalibrierwesen

- Bilderkennung, Texterkennung (OCR mit neuronalem Netz), …

Motivation

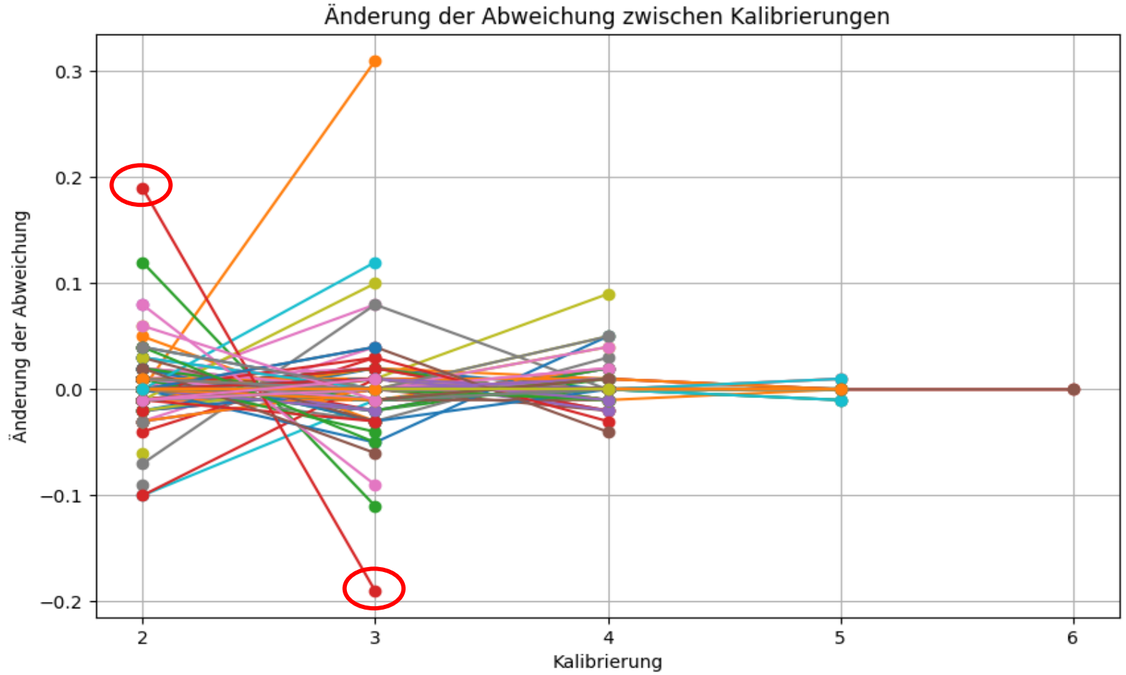
Validierung KI-Ergebnisse in der Qualitätssicherung

- Problematik von KI-Anwendungen
  - Halluzinieren, Blackbox, …
- Anwendung keine Vorhersage von Kalibrierergebnisse
- Fokus auf Prüfmittel-Verantwortlichen (Kunden)
  - „Risikobasierter Ansatz“ (DIN EN ISO/IEC 17025:2018)
  - Kalibrierintervall

Motivation

Abweichung Messwerte stabil

- Beobachtung: Abweichungen zwischen Messwerten statistisch stabil



- These: „Lassen abweichende Messwerte (Anomalien) Rückschlüsse über das Gerät zu?“
  - Anomalie Erkennung: Isolation Forest

Isolation Forest

Einleitung Isolation Forest

- Paper Liu et al. - Isolation Forest, 2008
  - Anomalie-Erkennung
  - Anwendung bei Kreditkartenbetrug
- Alternativen:
  - Z-Score
    - statistisches Maß, Anzahl Standardabweichungen vom Mittelwert
  - One-Class Support Vector Machine (SVM)
    - Lernt Entscheidungsgrenze um normale Datenpunkte
    - Schlecht bei großen Datensätzen

Isolation Forest

Fei Tony Liu, Kai Ming Ting
Gippsland School of Information Technology
Monash University, Victoria, Australia
{tony.liu},{kaiming.ting}@infotech.monash.edu.au

Zhi-Hua Zhou
National Key Laboratory
for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

Abstract

Most existing model-based approaches to anomaly detection construct a profile of normal instances, then identify instances that do not conform to the normal profile as anomalies. This paper proposes a fundamentally different model-based method that explicitly isolates anomalies instead of profiles normal points. To our best knowledge, the concept of isolation has not been explored in current literature. The use of isolation enables the proposed method, iForest, to exploit sub-sampling to an extent that is not feasible in existing methods, creating an algorithm which has a linear time complexity with a low constant and a low memory requirement. Our empirical evaluation shows that iForest performs favourably to ORCA, a near-linear time complexity distance-based method, LOF and Random Forests in terms of AUC and processing time, and especially in large data sets. iForest also works well in high dimensional problems which have a large number of irrelevant attributes, and in situations where training set does not contain any anomalies.

have different data char-

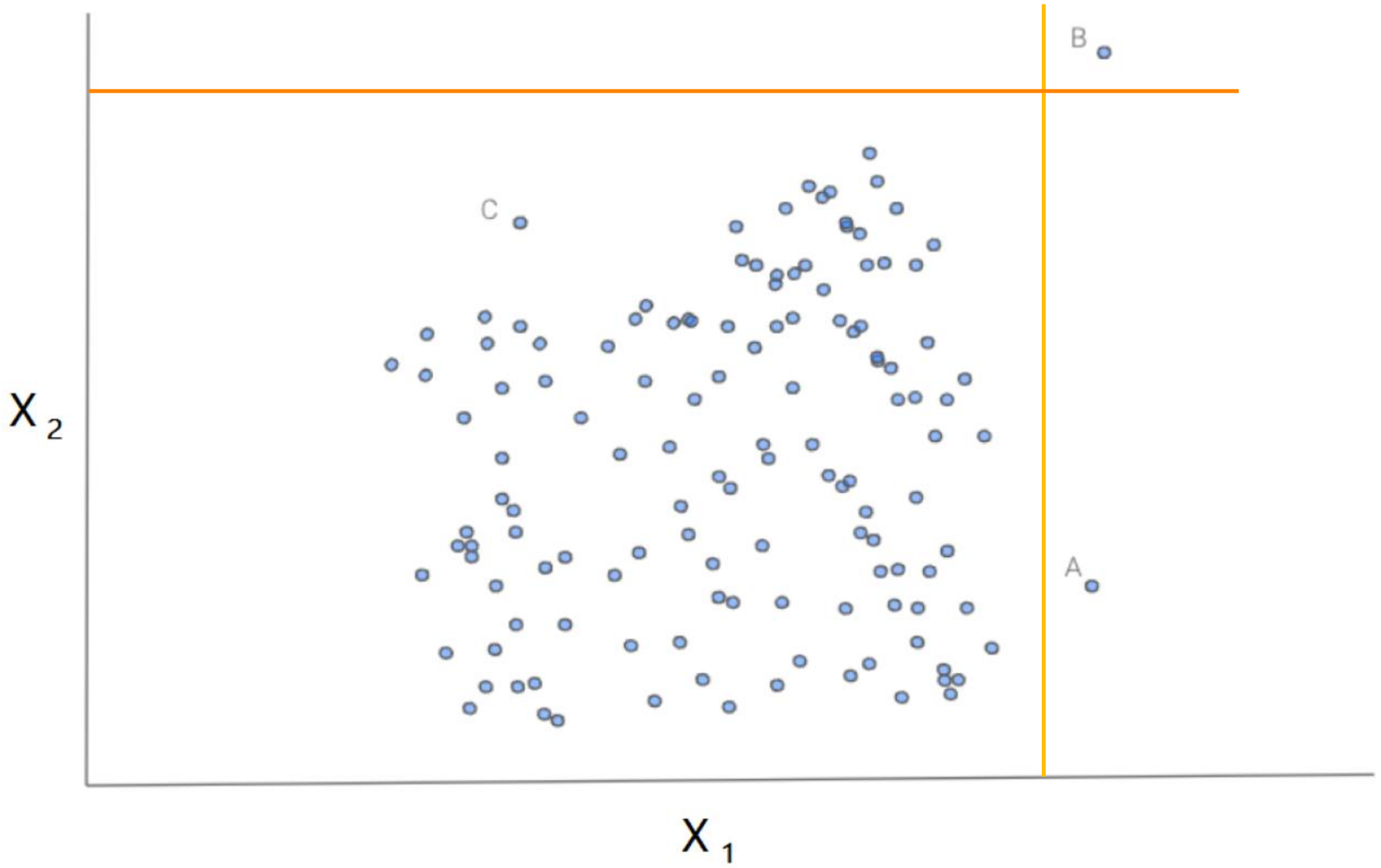
anomalies. Notable examples such as statistical methods [11], classification-based methods [1], and clustering-based methods [5] all use this general approach. Two major drawbacks of this approach are: (i) the anomaly detector is optimized to profile normal instances, but not optimized to detect anomalies—as a consequence, the results of anomaly detection might not be as good as expected, causing too many false alarms (having normal instances identified as anomalies) or too few anomalies being detected; (ii) many existing methods are constrained to low dimensional data and small data size because of their high computational complexity.

This paper proposes a different type of model-based method that explicitly isolates anomalies rather than profiles normal instances. To achieve this, our proposed method takes advantage of two anomalies' quantitative properties: i) they are the minority consisting of fewer instances and ii) they have attribute-values that are very different from those of normal instances. In other words, anomalies are 'few and different', which make them more susceptible to isolation than normal points. We show in this paper that a tree structure can be constructed effectively to isolate every single instance. Because of their susceptibility to isolation, anomalies are isolated closer to the root of the tree; whereas normal points are isolated at the deeper end of the tree. This isolation characteristic of tree forms the basis of our method

Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.

Isolation Forest

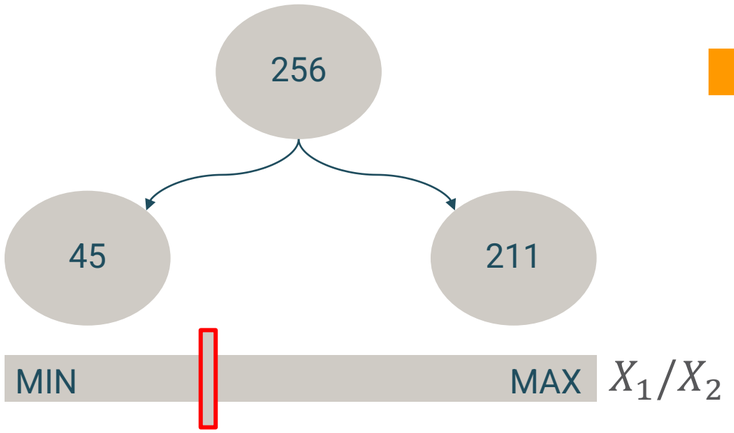
Beispiel Isolation Forest - Splits

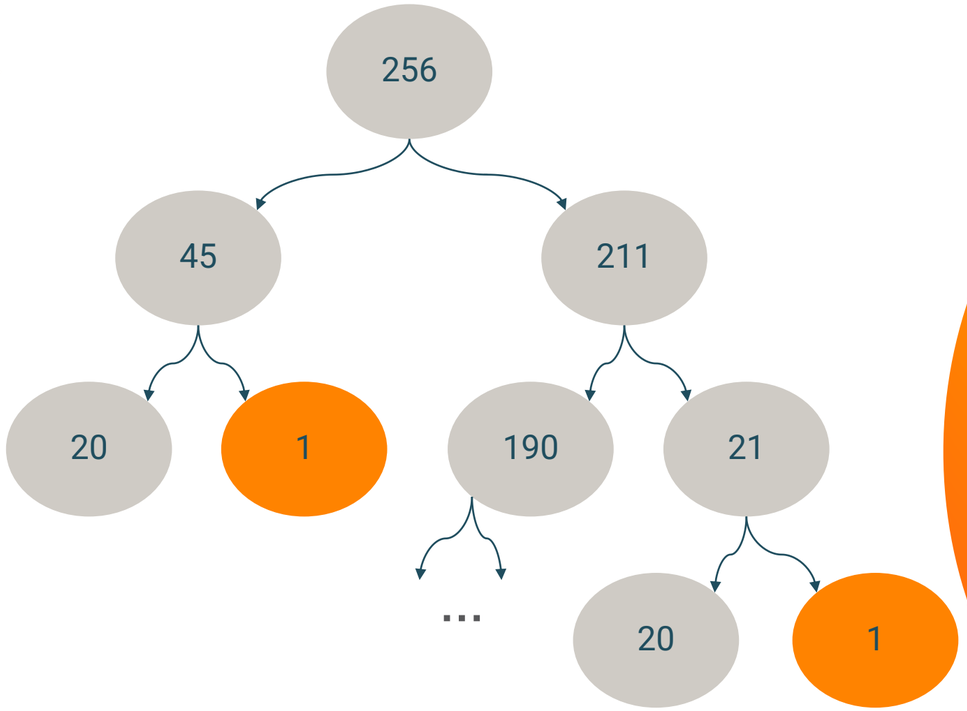


Isolation Forest Guide: Explanation and Python Implementation | DataCamp, Figure 2

Isolation Forest

Isolation Trees





- Baumstruktur
  - Zufällig Teilung der Features und Instanzen
  - Wiederholung: Aus vielen Bäumen wird ein Wald

Isolation Forest

Anomalie Score

- Mathematische Modellierung für *n* Instanzen

  $h(x)$ $-$ *Pfadlänge von Datenpunkt x*

  $E[h(x)]$ $-$ *Durchschnittliche Pfadlänge von allen Isolation Trees*

  $c(n)$ $-$ *Durchschnitt von h(x) in Abhängigkeit von n*

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right), \qquad mit\ H(i) = \sum_{k=1}^{i} \frac{1}{k},$$

$$s(x,n) = 2^{-\frac{E[h(x)]}{c(n)}} \in (0,1)$$

- $s(x,n) \xrightarrow{E[h(x)] > c(n)} 0$
- $s(x,n) \xrightarrow{E[h(x)] < c(n)} 1$

Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.

Isolation Forest

Vorteile/ Nachteile Isolation Forest

- Vorteile
  - Anwendung auf hohe Dimensionen
  - Keine Voranalyse der Daten/ Features nötig
  - Unüberwacht und Lineare Zeitkomplexität
    - In großen Datensätzen Anomalie-Erkennung möglich ohne vorherige Informationen
- Nachteile:
  - Anwendung auf kleine Datensätze
  - Nur Anomalie-Erkennung → Keine qualitative Aussage über Anomalie möglich

Isolation Forest

Implementierung Isolation Forest

- Open Source
  - Python-Bibliothek *scikit-learn*
- Variablen
  - n_estimators
    - Anzahl Isolation Trees → Standardwert: 100
  - max_samples
    - Trainingsdatensatz jedes Isolation Tree → Standardwert: 256
  - Kontamination
    - Anteil Equipments außerhalb der Toleranz

[https://github.com/scikit-learn/scikit-learn]

```
clf = IsolationForest(
    n_estimators=100,
    max_samples=256,
    contamination=contamination_var/factor_contamination
)
```

Datenanalyse

Machine Learning – Features

- Datenbestand für einzelnen Gerätetyp
- Feature Messgröße
  - (Messgröße, Referenzwert, Einheit, Messbedingung, Einheit)

```
('ACU', 30.0, 'V', 50.0, 'Hz')
('ACU', 30.0, 'V', 500.0, 'Hz')
('ACU', 60.0, 'mV', 50.0, 'Hz')
('C', 10.0, 'µF', '-', '-')
('C', 100.0, 'µF', '-', '-')
('DCI', 0.0, 'mA', '-', '-')
('DCR', 0.0, 'Ohm', '-', '-')
('DCU', -6.0, 'V', '-', '-')
```

- Feature Messgröße sortiert nach Anzahl Messpunkten im Datensatz
- Werte
  - Abweichung
  - Änderung der Abweichung
- Entfernen aller Messpunkte außerhalb der Toleranz

Datenanalyse

Skizze Datenaufbereitung

Equipment-nummer	Kalibrier-datum	Zertifikat	Anzahl Kalibrierungen	Feature Messgröße 1	…	Feature Messgröße N
EQ 1	dd.mm.yyyy	Zert. 1A	1	Wert (1,1)	…	Wert (1,N)
EQ 2	dd.mm.yyyy	Zert. 2A	1	Wert (2,1)	…	Wert (2,N)
EQ 2	dd.mm.yyyy	Zert. 2B	2	Wert (3,1)	…	Wert (3,N)
…	…	…	…	…	…	…
EQ X	dd.mm.yyyy	Zert. XE	5	Wert (K-1,1)	…	Wert (K-1,N)
EQ X	dd.mm.yyyy	Zert. Z	6	Wert (K,1)	…	Wert (K,N)

- Wert (X,Y) nicht vorhanden → Mittelwert der jeweiligen Feature Messgröße

Datenanalyse

Datensatz A

- Digitalmultimeter Typ A

Messpunkte	Equipments	Equipments „Fail“	Anzahl Feature Messgröße
539.772	3727	55	548

- Viele Messpunkte
- Geringer Anteil Equipments außerhalb der Toleranz

Methodik

Identifizierung Equipments mit Kalibrierergebnissen außerhalb der Toleranz

Datensatz A

Isolation Forest: Datensatz ohne Messpunkte außerhalb Toleranz

Datensatz A

Parameter bei der Equipment-Erkennung:

- Anzahl Features Messgröße
- Werte: Abweichung/ Änderung der Abweichung
- Kontamination: Verhältnis Equipments außerhalb der Toleranz

Datensatz A

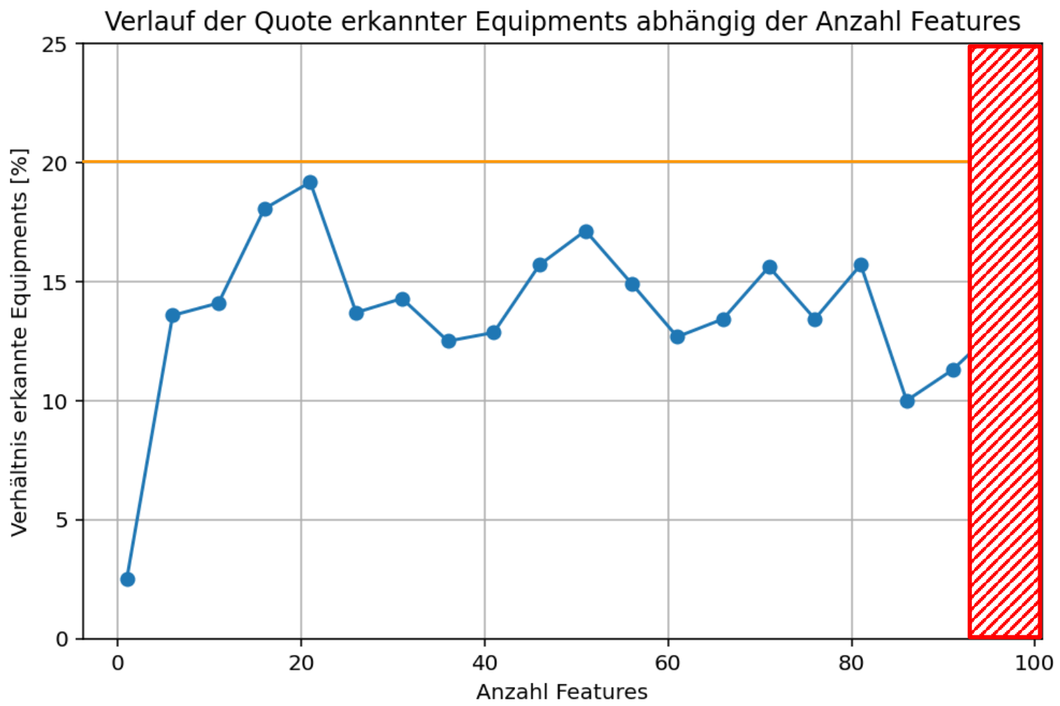
Identifikation: Equipments mit Messpunkt außerhalb der Toleranz

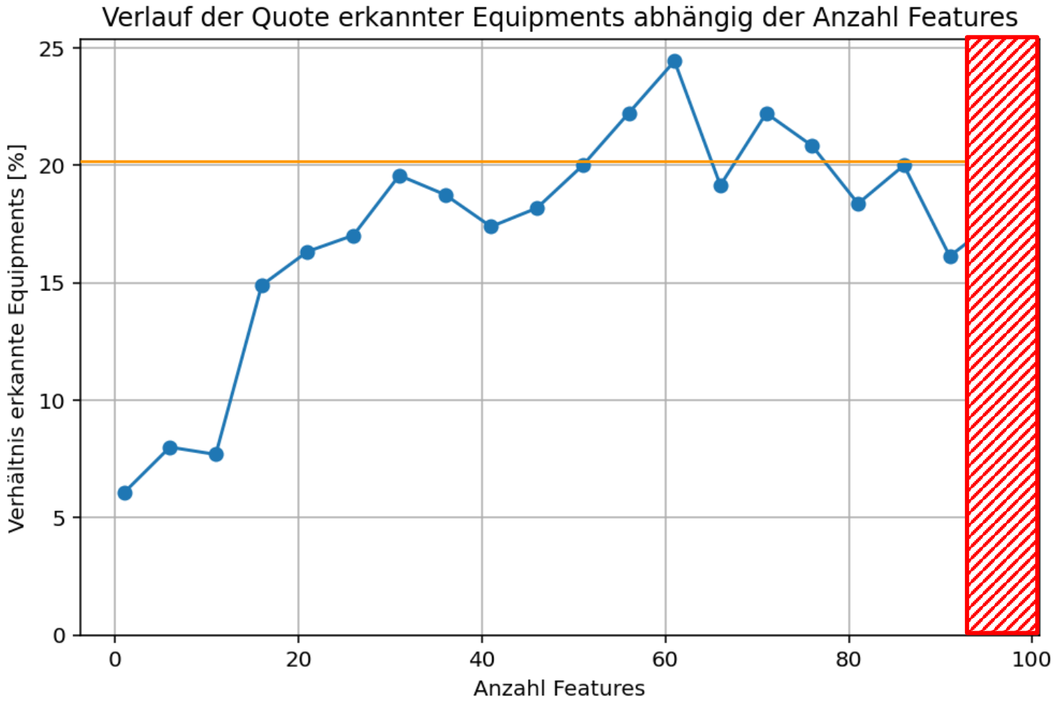
Datensatz B

Validieren der Ergebnisse mit den gleichen Parametern

Ergebnisse

Datensatz A – Anzahl Features Messgröße





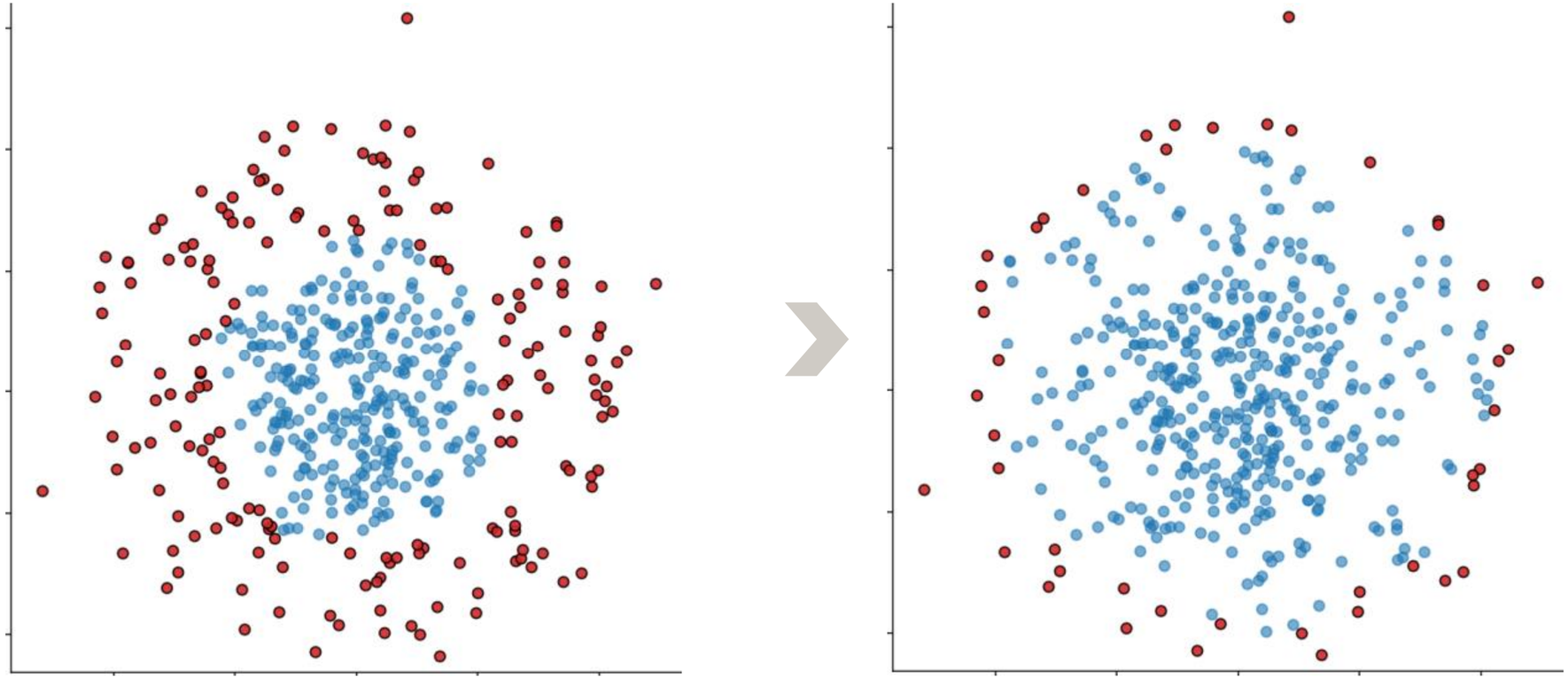
- Abweichung
  - Tendenz unter 20%

- Änderung der Abweichung
  - Bei ~50-75 Features >20%

Ergebnisse

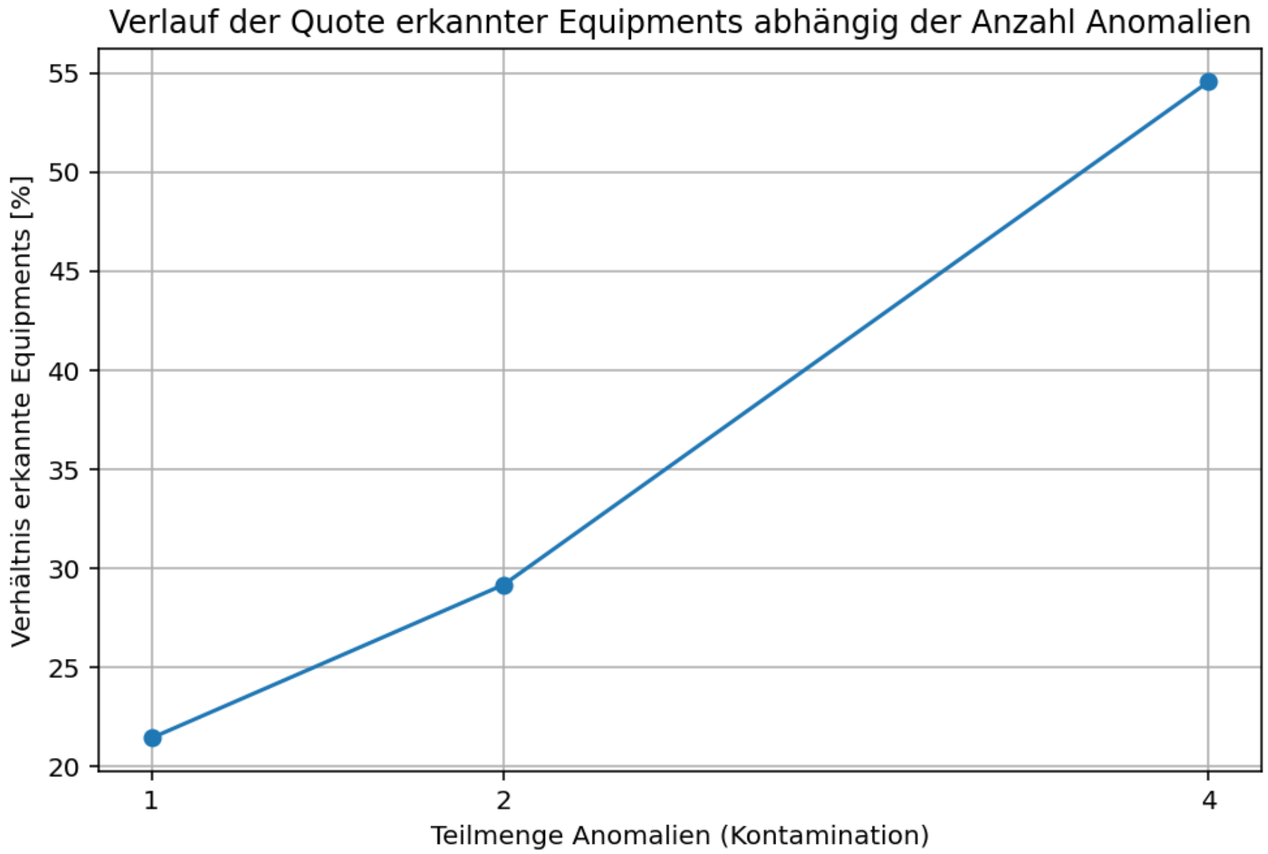
Ergebnisse – Kontamination variieren

- Anteil der Kontamination wählen
  - Geringere Anzahl Anomalien (Höherer Anomalie Score)



Ergebnisse

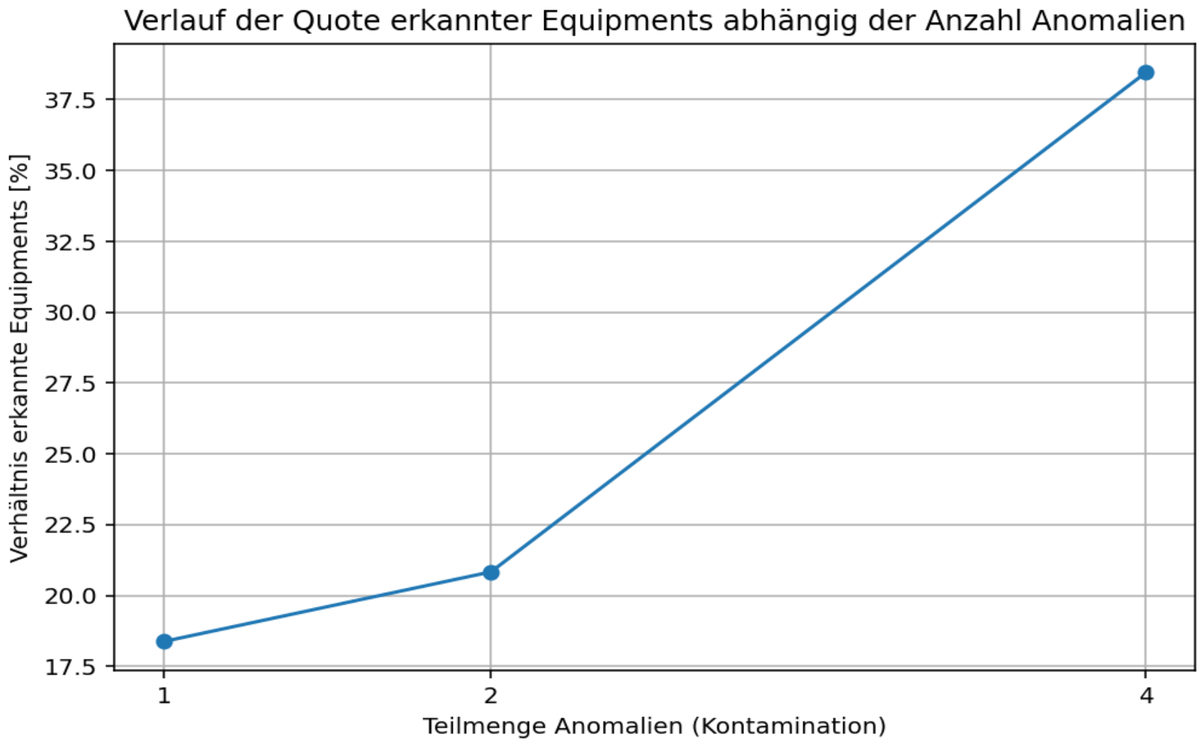
Anteil Kontamination



- Anteil Anomalien
  - Datensatz A

Ergebnisse

Anteil Kontamination – Validierung



- Anteil Anomalien
  - Datensatz B (Änderung Abweichung, 60 Features Messgröße)

Messpunkte	Equipments	Equipments „Fail“	Anzahl Feature Messgröße
502.745	3366	51	517

Datenanalyse

Datensatz C

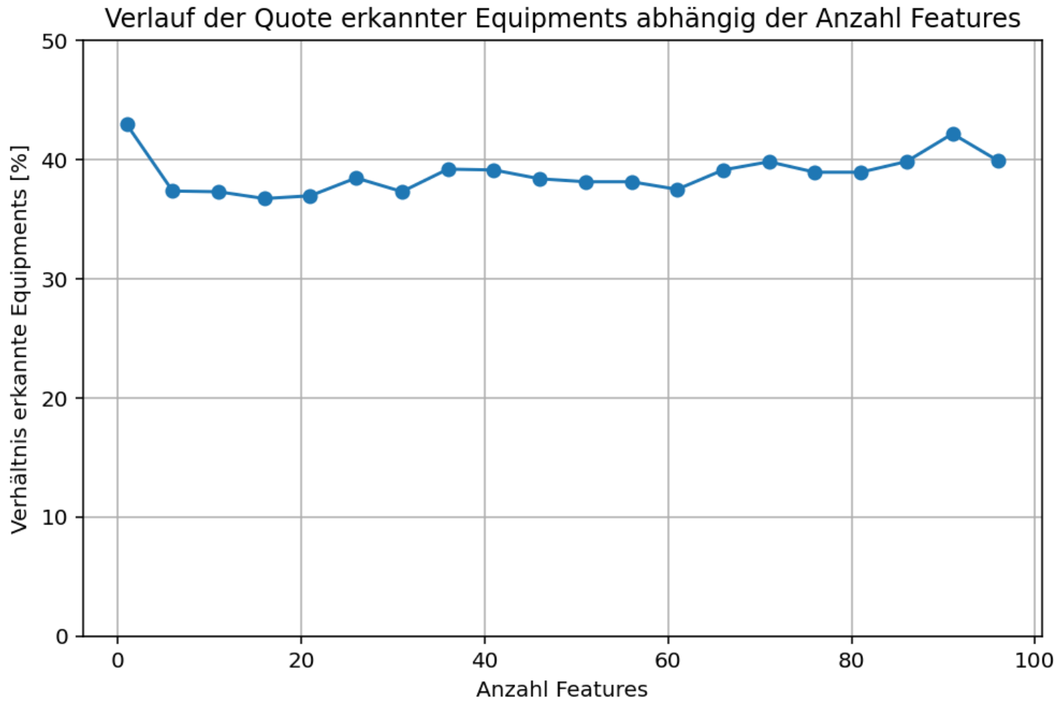
- Poweranalyser Typ C

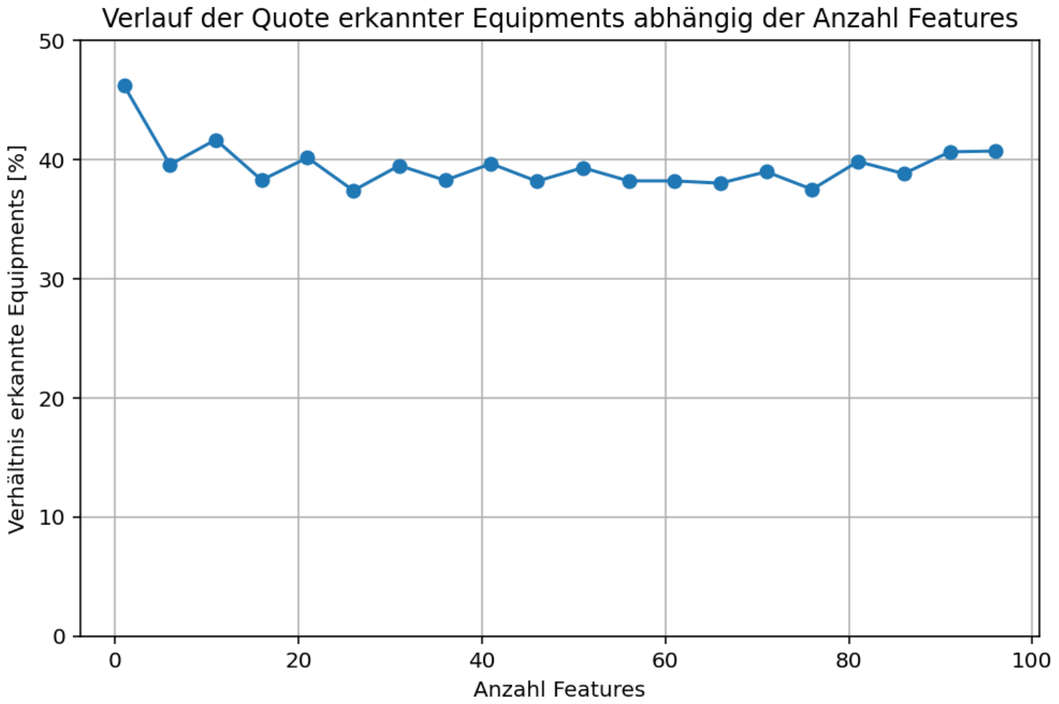
Messpunkte	Equipments	Equipments „Fail“	Anzahl Feature Messgröße
114.590	302	94	1215

- Anteil Equipments außerhalb der Toleranz sehr hoch
- Hohe Anzahl Feature Messgröße

Ergebnisse

Datensatz C





- Abweichung
  - Datensatz C
- Änderung der Abweichung
  - Datensatz C

Ergebnisse

Ergebnisse – Datensatz „viele Messpunkte außerhalb“

- Datensatz C
  - Sehr hohe Quote Equipments außerhalb
    - 307 von 985 Kalibrierungen (31,2%)
  - Nicht gut geeignet

Ausblick

Weiteres Vorgehen

- Fazit
  - Zusammenhang zwischen Messpunkt-Anomalien und Equipments mit Messpunkten außerhalb der Toleranz
- Weitere Messgrößen/ Datensätze
- Möglichkeit Gerätetypen unterschiedlicher Hersteller zu vergleichen
  - z.B. Anzahl Anomalien
- Einsatz im Kalibrierprozess
  - Methodik definieren
    - Gerätetyp, Art von Prüfmitteln, …
  - Nach Kalibrierung mit Abweichung innerhalb der Toleranz: Kalibrierergebnisse Anomalie?

Vielen Dank für die Aufmerksamkeit!

Fragen?

Philipp Alfter
Innovation
Phone: +49 7661 90901 8169
E-Mail: PAlfter@testotis.de