

Anomaly Detection zur Erkennung fehlerhafter Ladeprozesse in Kundendatenbanken

Henryk Borzymowski und Selina Reinicke

Statistisches Consulting
Projektpartner: DYMATRIX CONSULTING GROUP GmbH
Betreuung: Dr. Fabian Scheipl

München, 11. April 2018



Gliederung

Einleitung

Projektpartner und Aufgabenstellung

Datengrundlage

Anomaly Detection

Anomaly Detection Methoden

Bewertung von unsupervised Anomaly Detection Methoden

Probleme

Übersicht zu erhaltenen Daten

Produkt für Projektpartner und Zusammenfassung

Gliederung

Einleitung

Projektpartner und Aufgabenstellung

Datengrundlage

Anomaly Detection

Anomaly Detection Methoden

Bewertung von unsupervised Anomaly Detection Methoden

Probleme

Übersicht zu erhaltenen Daten

Produkt für Projektpartner und Zusammenfassung

Projektpartner und Aufgabenstellung

- ▶ DYMATRIX ist Beratungsunternehmen und Dienstleister für digitales Marketing
- ▶ Cross-Channel-Marketing basiert auf einheitlichen Kundendaten
- ▶ Überprüfung der Ladeprozesse durch Anomaly Detection

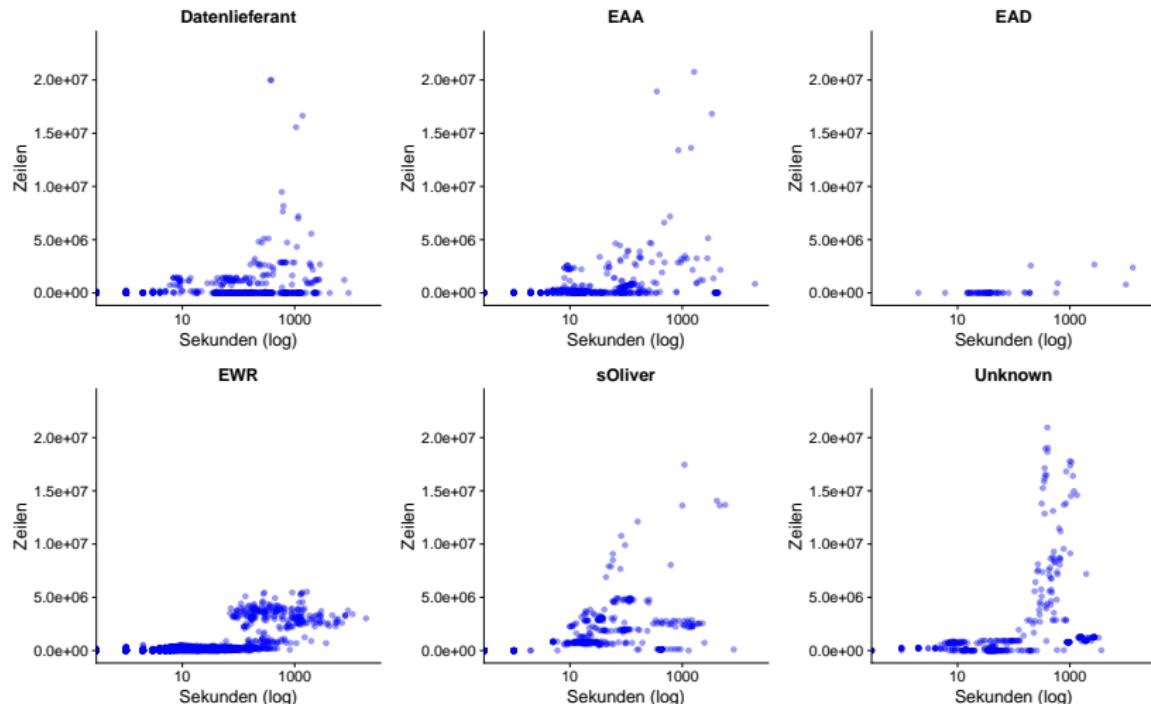
Datengrundlage

- ▶ Log-Dateien der Ladeprozesse
- ▶ Keine Labels
- ▶ Zeitraum: März/April - November 2017

Startzeit	...	Sekunden	Zeilen	Job	Package	...
...						
...						
<i>9.679 Zeilen</i>						

- ▶ 2 Variablen werden als relevant für Anomaly Detection identifiziert
 - ▶ „Sekunden“ für Dauer,
 - ▶ „Zeilen“ für Umfang des Ladeprozesses

Übersicht der Jobs



Gliederung

Einleitung

Projektpartner und Aufgabenstellung

Datengrundlage

Anomaly Detection

Anomaly Detection Methoden

Bewertung von unsupervised Anomaly Detection Methoden

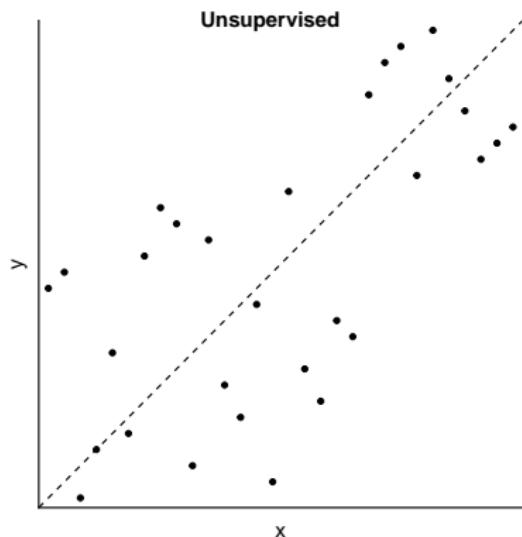
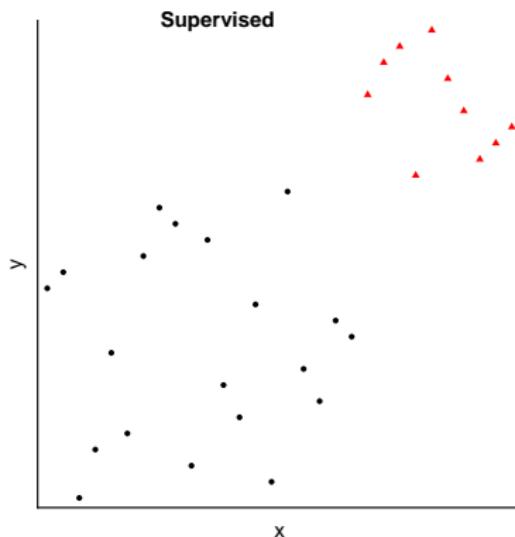
Probleme

Übersicht zu erhaltenen Daten

Produkt für Projektpartner und Zusammenfassung

Was ist eine Anomalie?

- ▶ Observationen, die „anders und selten“ sind
- ▶ Strukturwissen für Anomaly Detection-Methode hilfreich



Isolation forest (iForest) Methode

- ▶ Methode zur Erkennung von isolierten Datenpunkten
- ▶ Isolierte Punkte können als Anomalien bezeichnet werden
- ▶ Ein iForest entsteht als Ensemble von zufälligen Entscheidungsbäumen

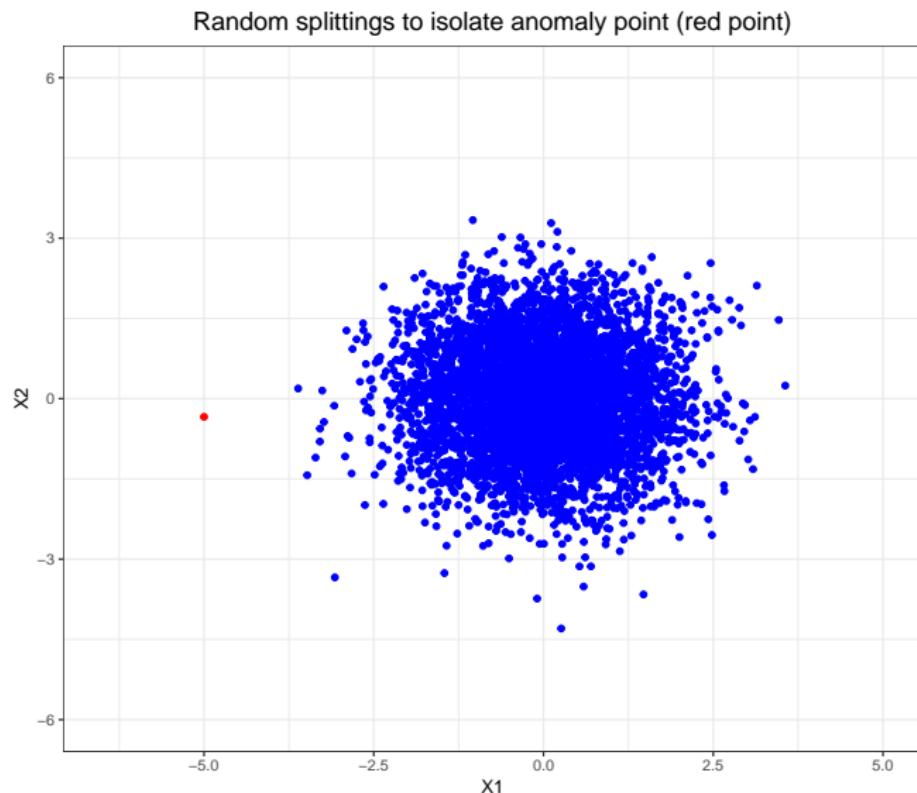


(a) Seehaus

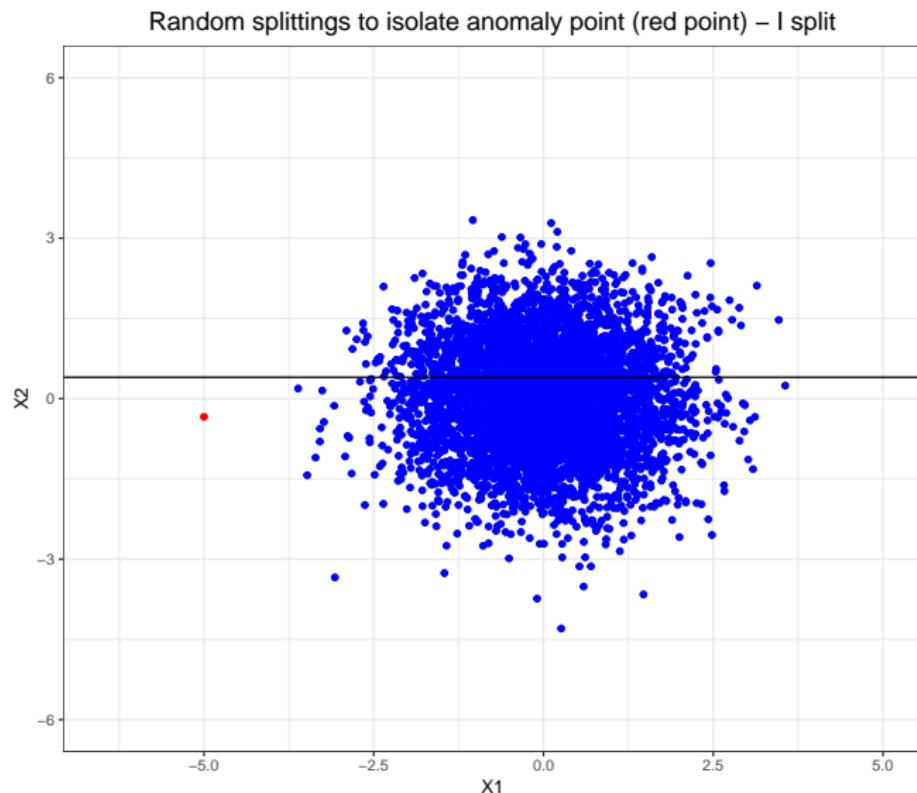


(b) Reihenhaus

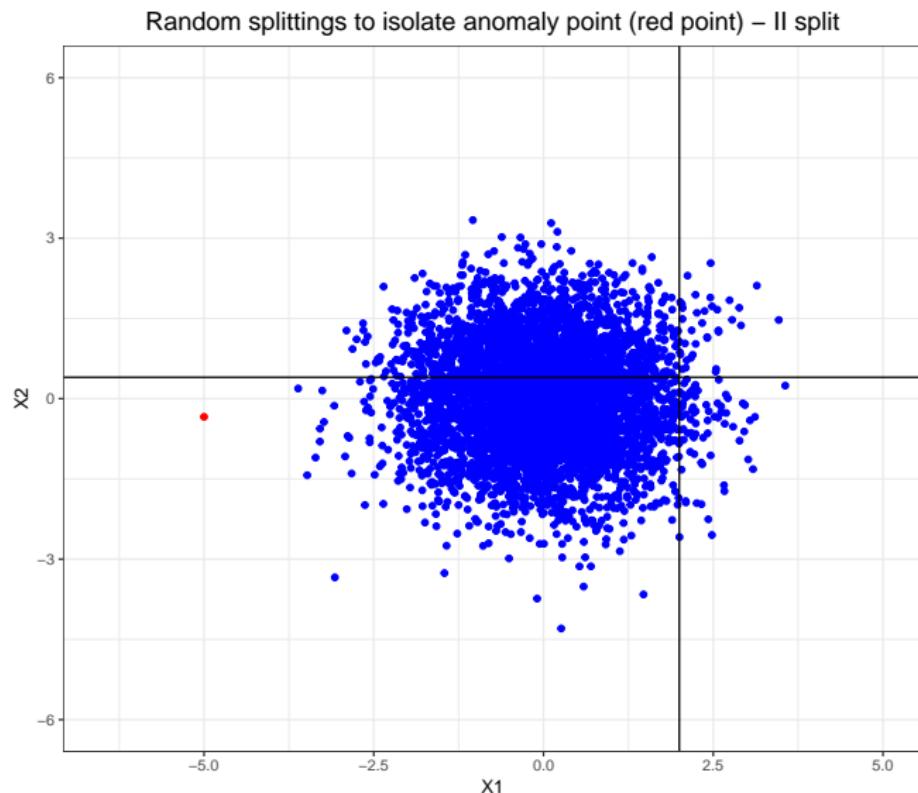
Random splitting für einen anomalen Datenpunkt



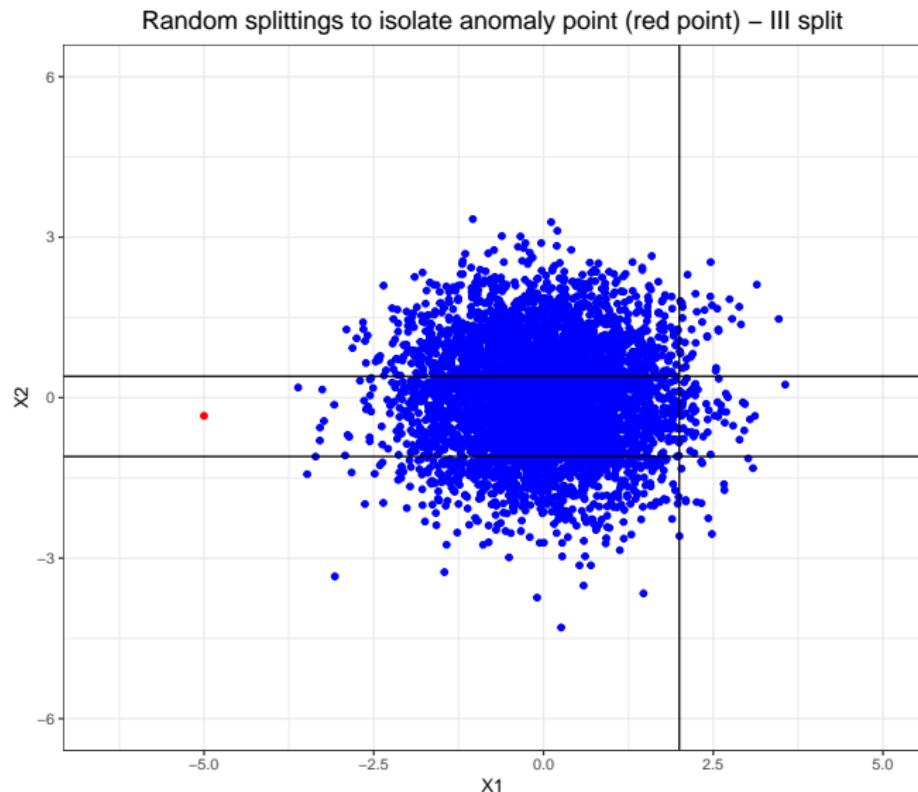
Random splitting für einen anomalen Datenpunkt



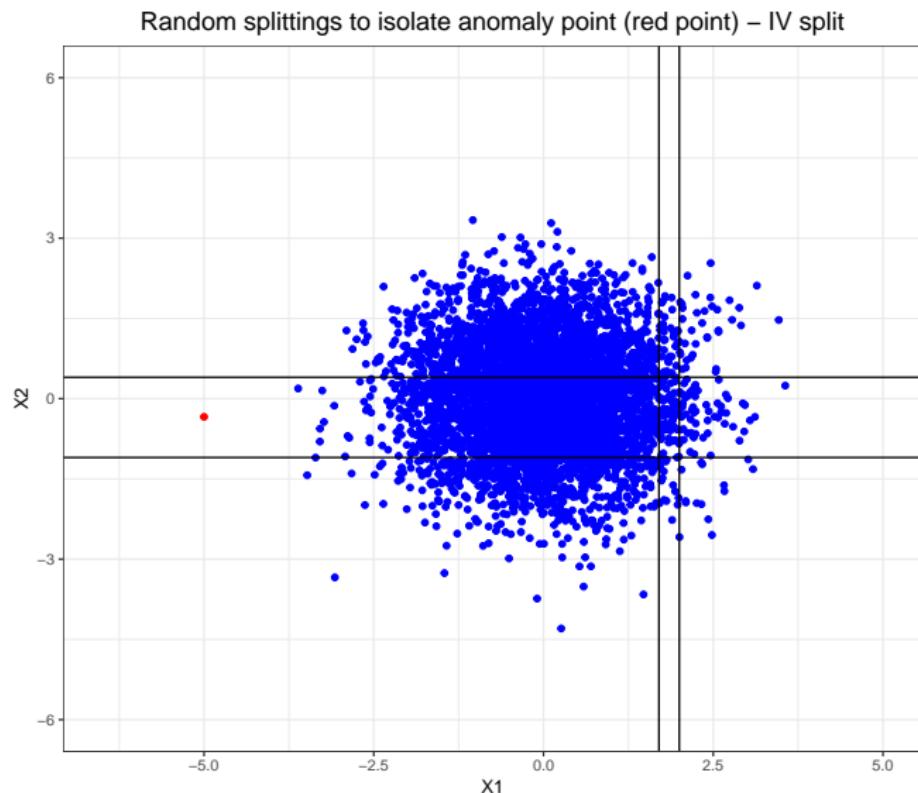
Random splitting für einen anomalen Datenpunkt



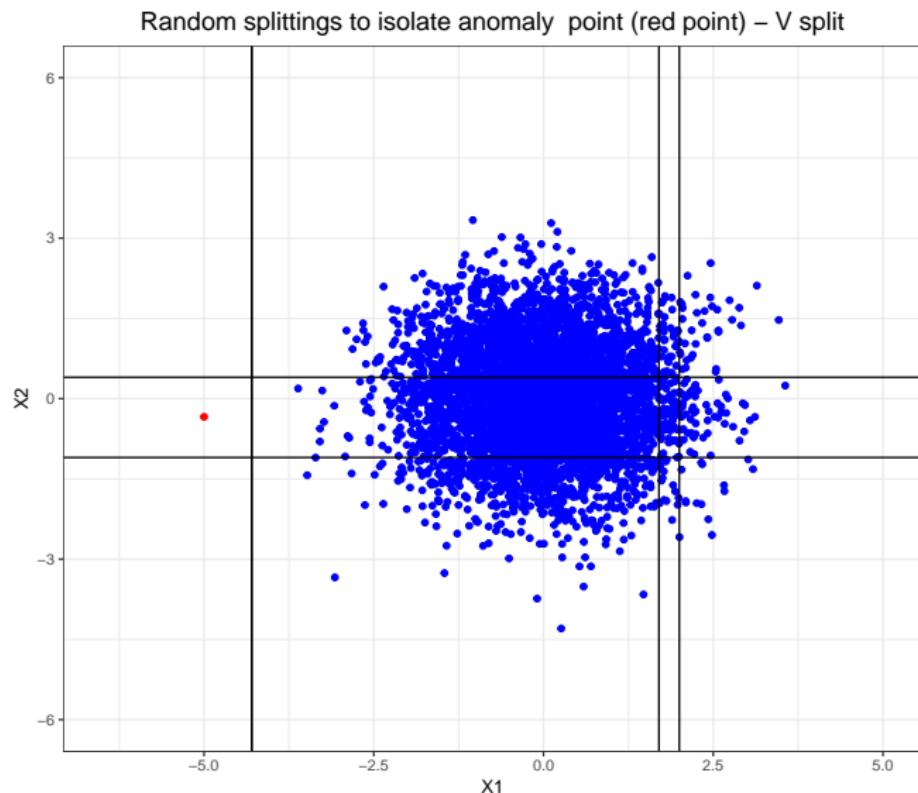
Random splitting für einen anomalen Datenpunkt



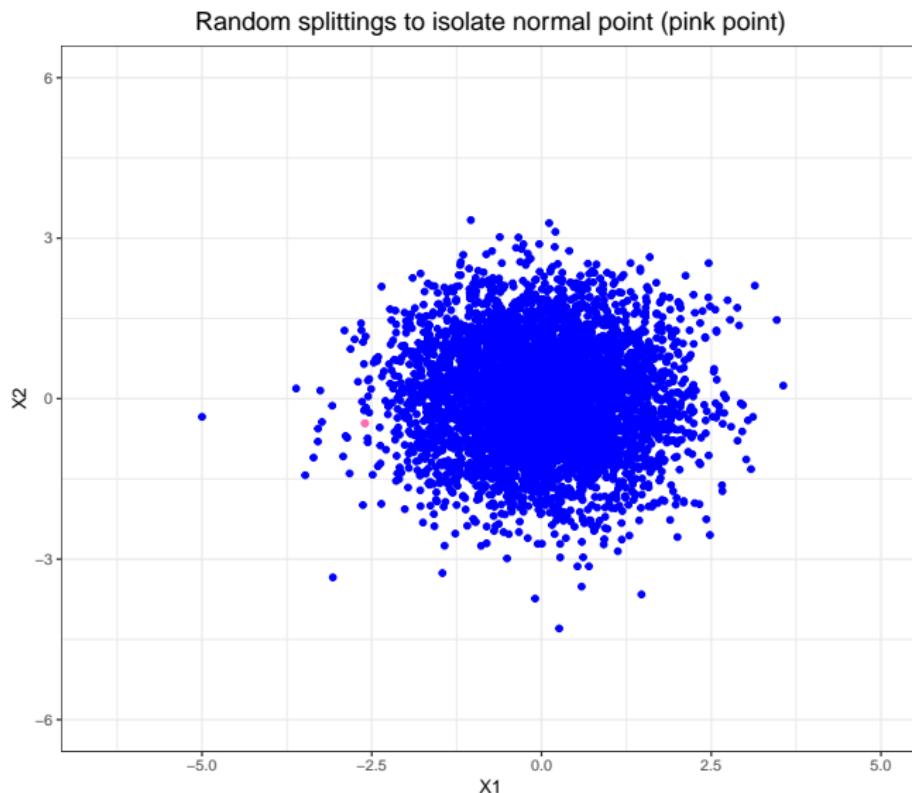
Random splitting für einen anomalen Datenpunkt



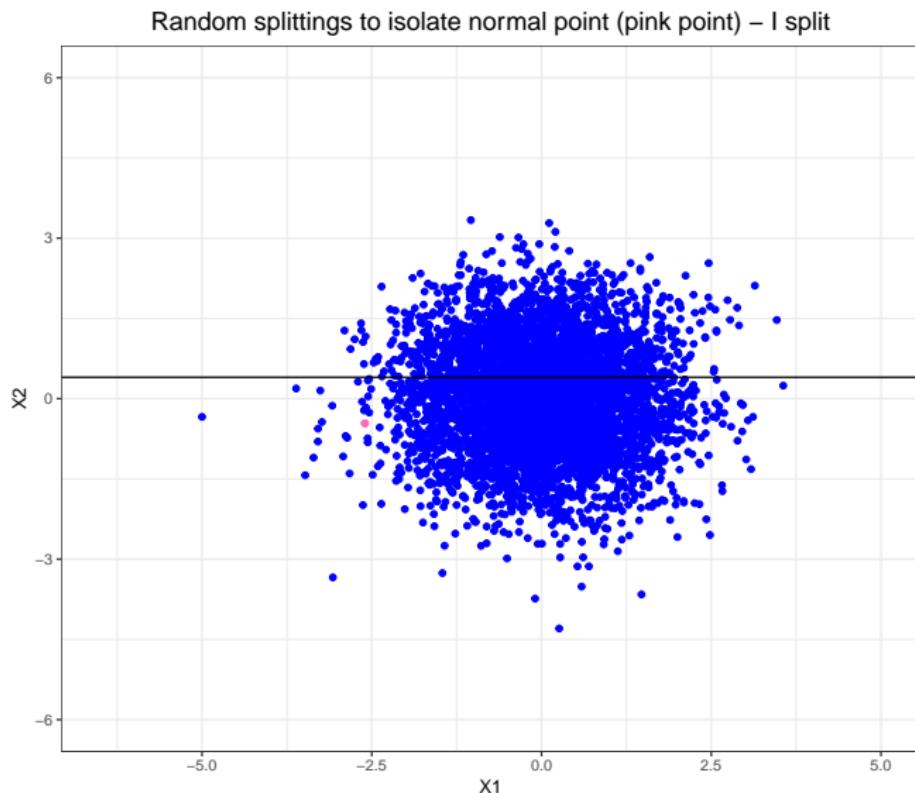
Random splitting für einen anomalen Datenpunkt



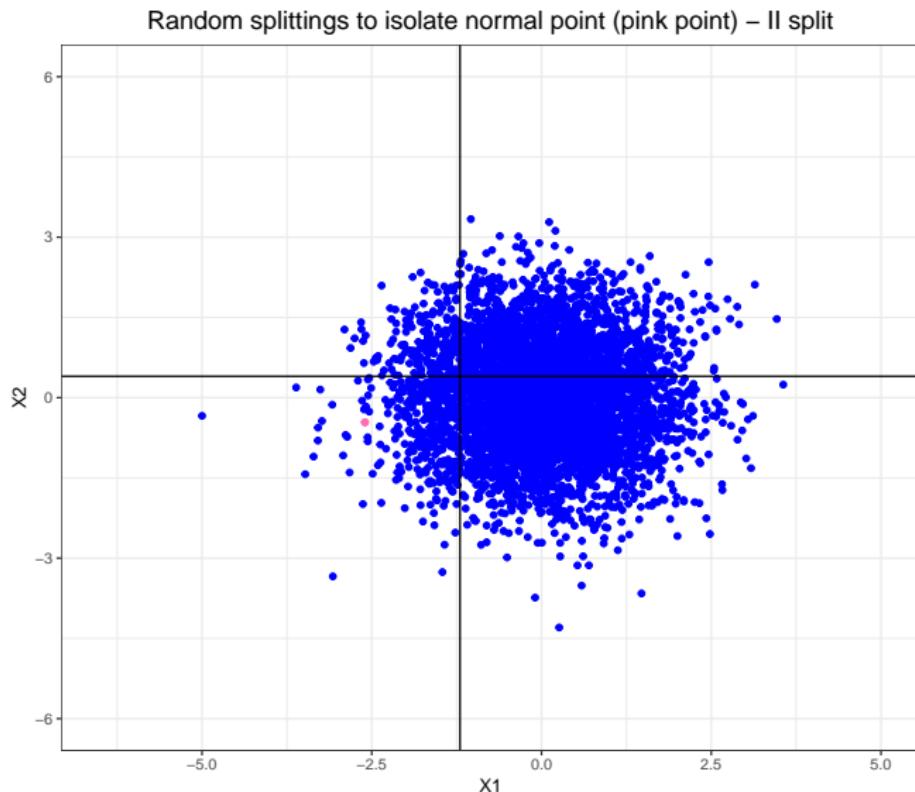
Random splitting für einen normalen Datenpunkt



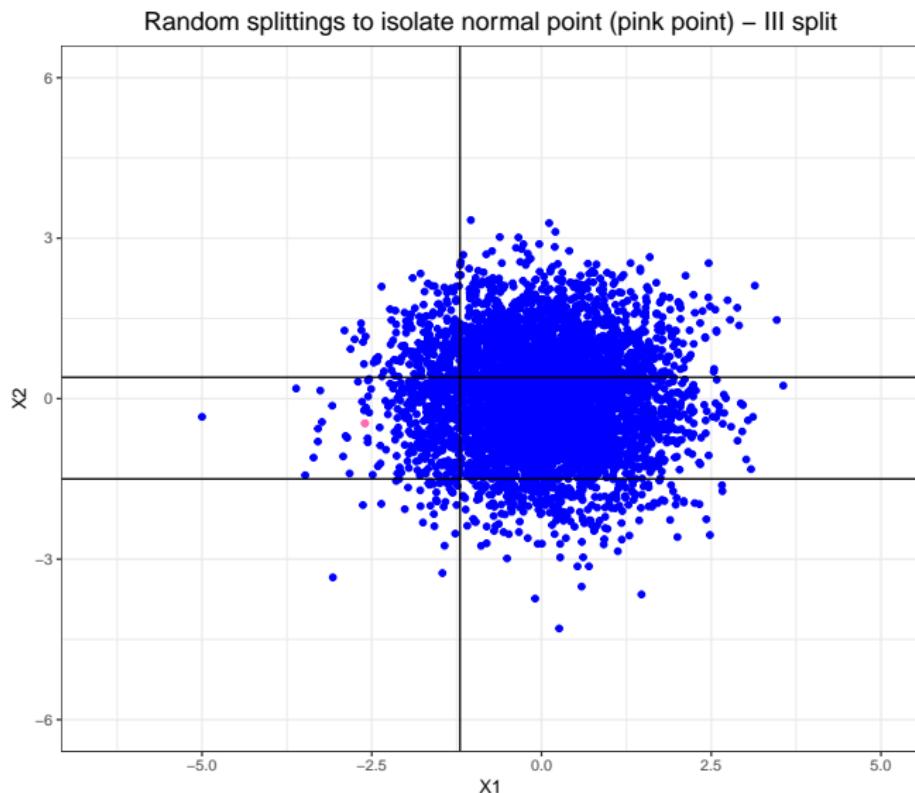
Random splitting für einen normalen Datenpunkt



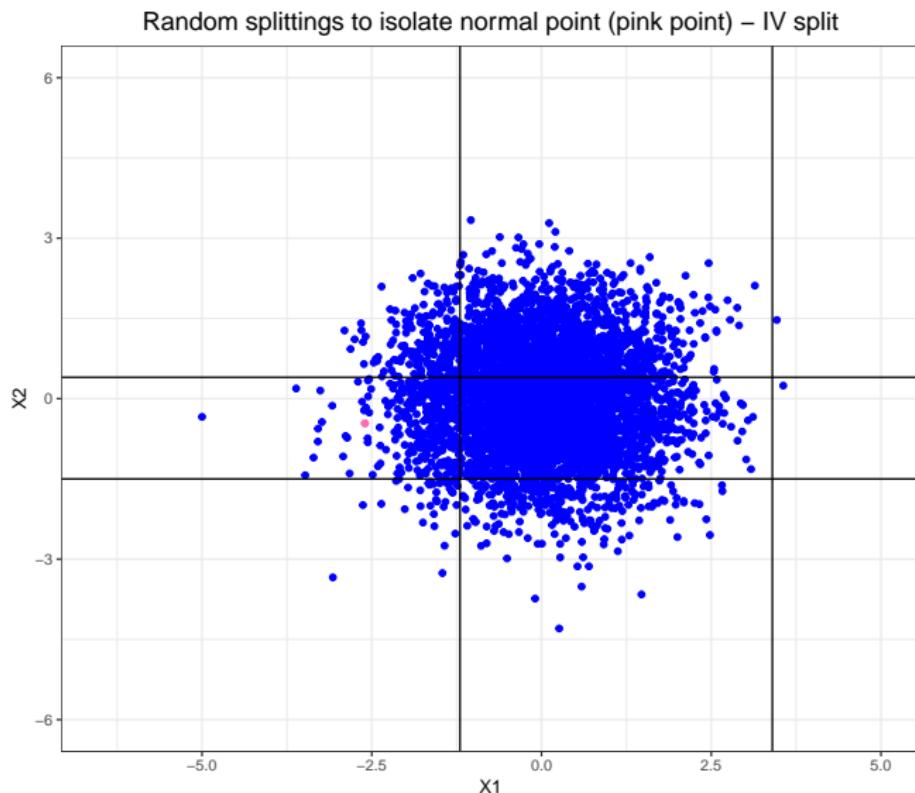
Random splitting für einen normalen Datenpunkt



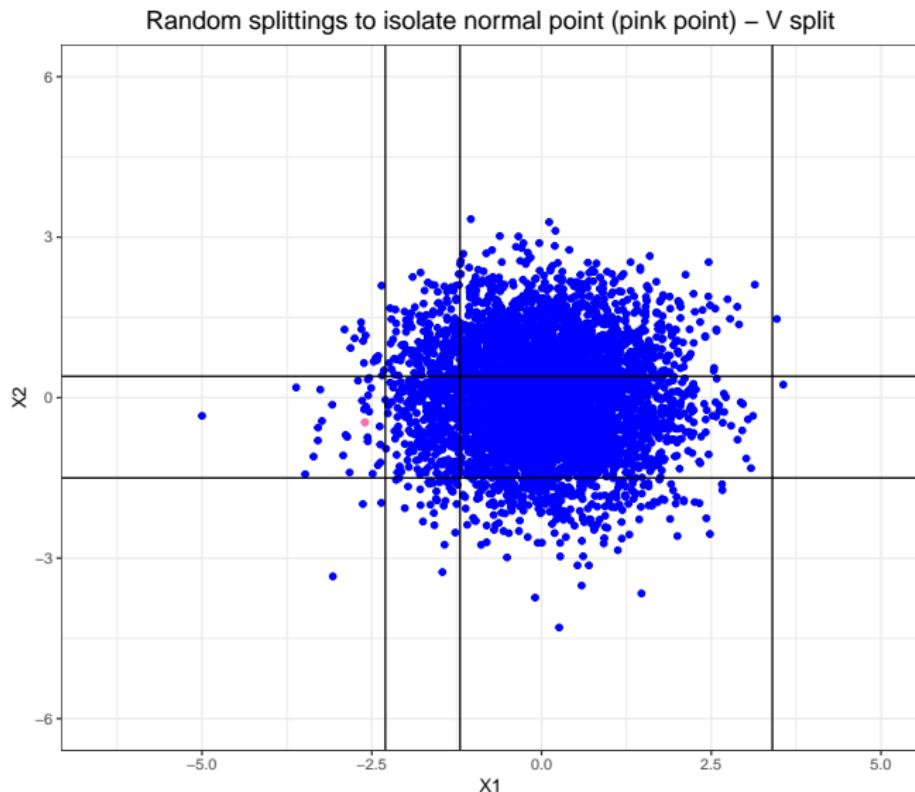
Random splitting für einen normalen Datenpunkt



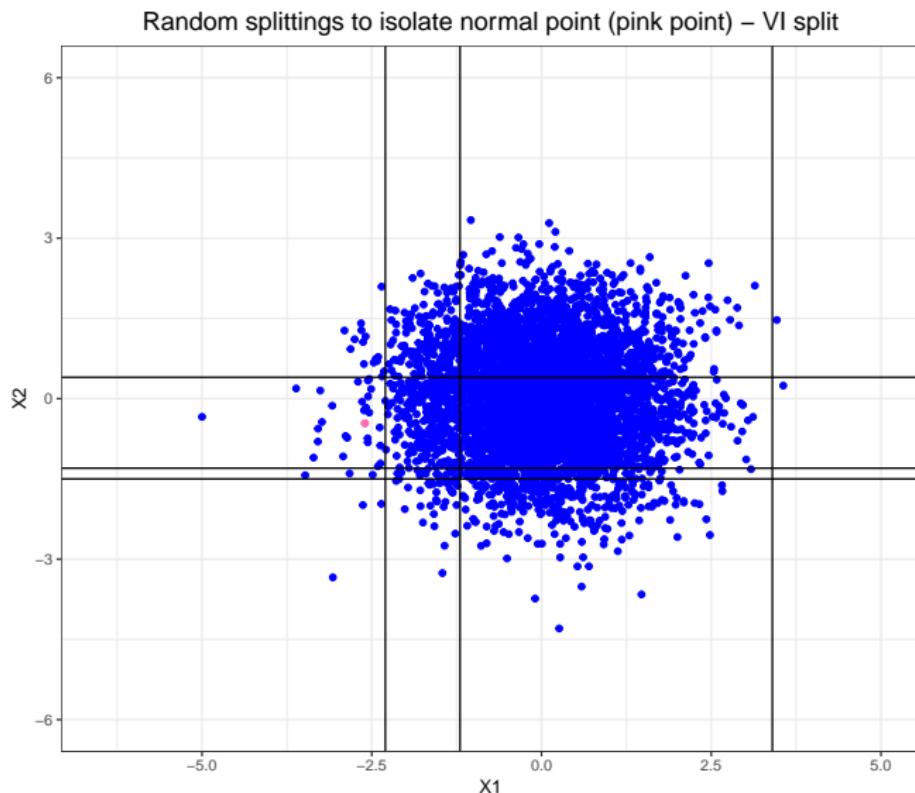
Random splitting für einen normalen Datenpunkt



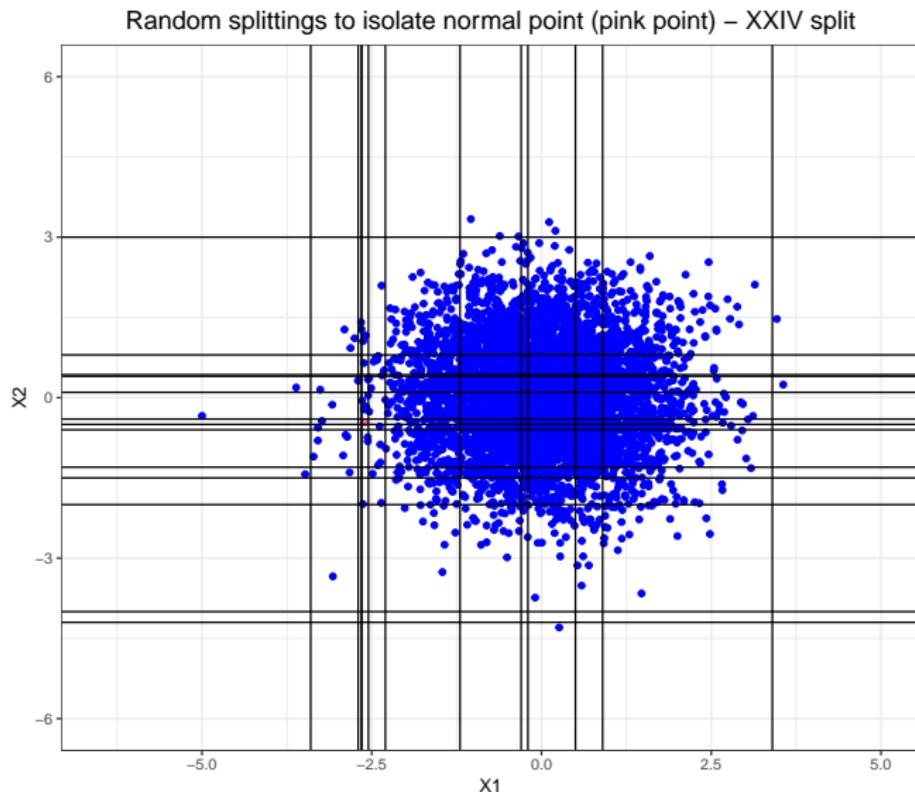
Random splitting für einen normalen Datenpunkt



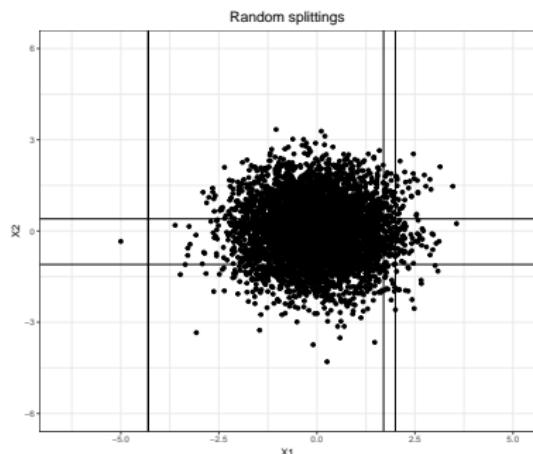
Random splitting für einen normalen Datenpunkt



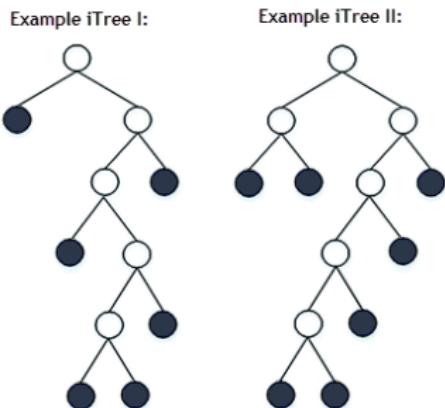
Random splitting für einen normalen Datenpunkt



Darstellung von splits als Baumstruktur (iTree)



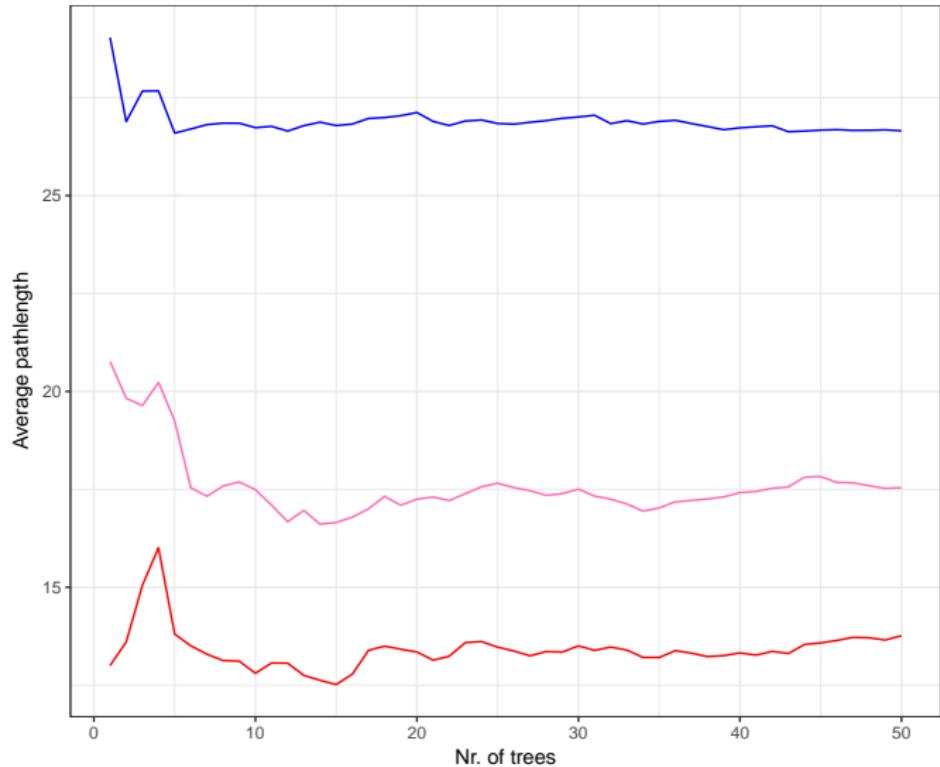
(c) Random splitting points



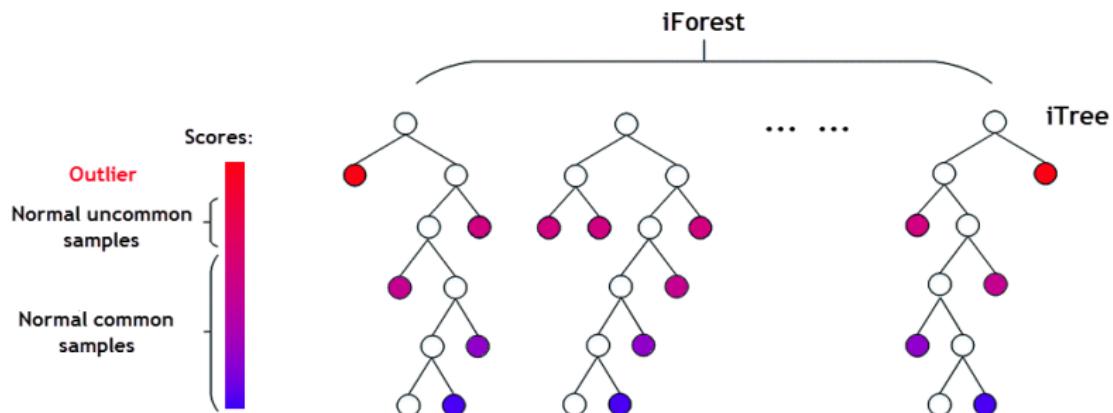
(d) Tree structure

Übergang von split Punkten zu einer Baumstruktur

Konvergenz der Pfadlängen



Die iForest Methode



Scores iForest berechnen

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

- ▶ $E(h(x))$ ist die durchschnittliche Länge der Äste, von einer Gruppe von iTrees, für die Observation x
- ▶ $c(n)$ ist die durchschnittliche Länge aller Äste in einem Baum

Scores iForest berechnen

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

- ▶ $E(h(x))$ ist die durchschnittliche Länge der Äste, von einer Gruppe von iTrees, für die Observation x
- ▶ $c(n)$ ist die durchschnittliche Länge aller Äste in einem Baum

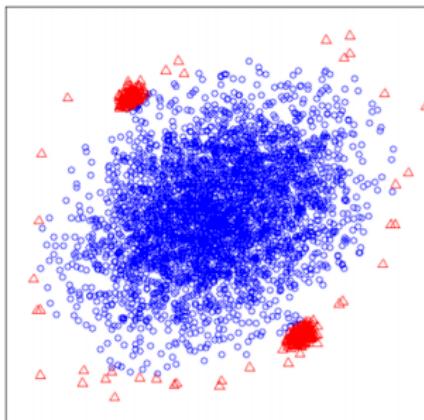
Tabelle: Anomaly Scores und Pfadlängen.

Pfadlängen	Score	Anomalie?
$E(h(x)) \rightarrow c(n)$	$s \rightarrow 0.5$	Eher keine Anomalie
$E(h(x)) \rightarrow 0$	$s \rightarrow 1$	Wahrscheinlich eine Anomalie
$E(h(x)) \rightarrow n - 1$	$s \rightarrow 0$	Keine Anomalie

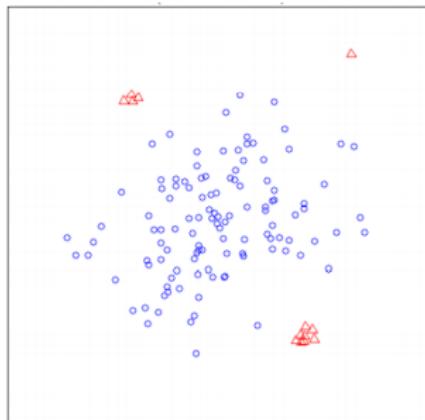
Quelle : Fei Tony Liu, Kai Ming Ting Zhi-Hua Zhou (2013) *Isolation Forest*

Swamping und masking

- ▶ Swamping: Fehlerhafte Einstufung als Anomalie
- ▶ Masking: Zu dichte Anomalien heben sich gegenseitig auf



(a) Original sample



(b) Sub-sample

Quelle : Fei Tony Liu, Kai Ming Ting Zhi-Hua Zhou (2013) *Isolation Forest*

Vor- und Nachteile der iForest Methode

Vorteile:

- ▶ Funktioniert gut für hohe Dimensionen
- ▶ Komputational effizient
- ▶ Braucht keine Labels
- ▶ Funktioniert für Datensätze die Anomalien enthalten

Vor- und Nachteile der iForest Methode

Vorteile:

- ▶ Funktioniert gut für hohe Dimensionen
- ▶ Komputational effizient
- ▶ Braucht keine Labels
- ▶ Funktioniert für Datensätze die Anomalien enthalten

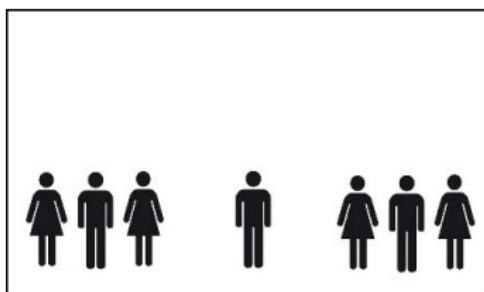
Nachteile:

- ▶ Relativ neu
- ▶ Keine Interpretation möglich

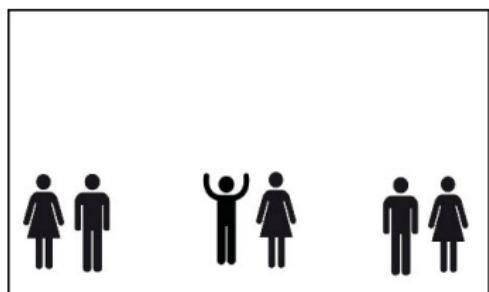
Local Outlier Factor (LOF)

Eine Methode zur Erkennung von dichtebasierten Anomalien:

- ▶ Berechnung von Dichten anhand von Distanzen zwischen Datenpunkten
- ▶ Vergleich der Dichten innerhalb der Nachbarschaft
- ▶ Berechnung von Scores (LOF) je Datenpunkt
- ▶ Parameter: Anzahl der Nachbarn k

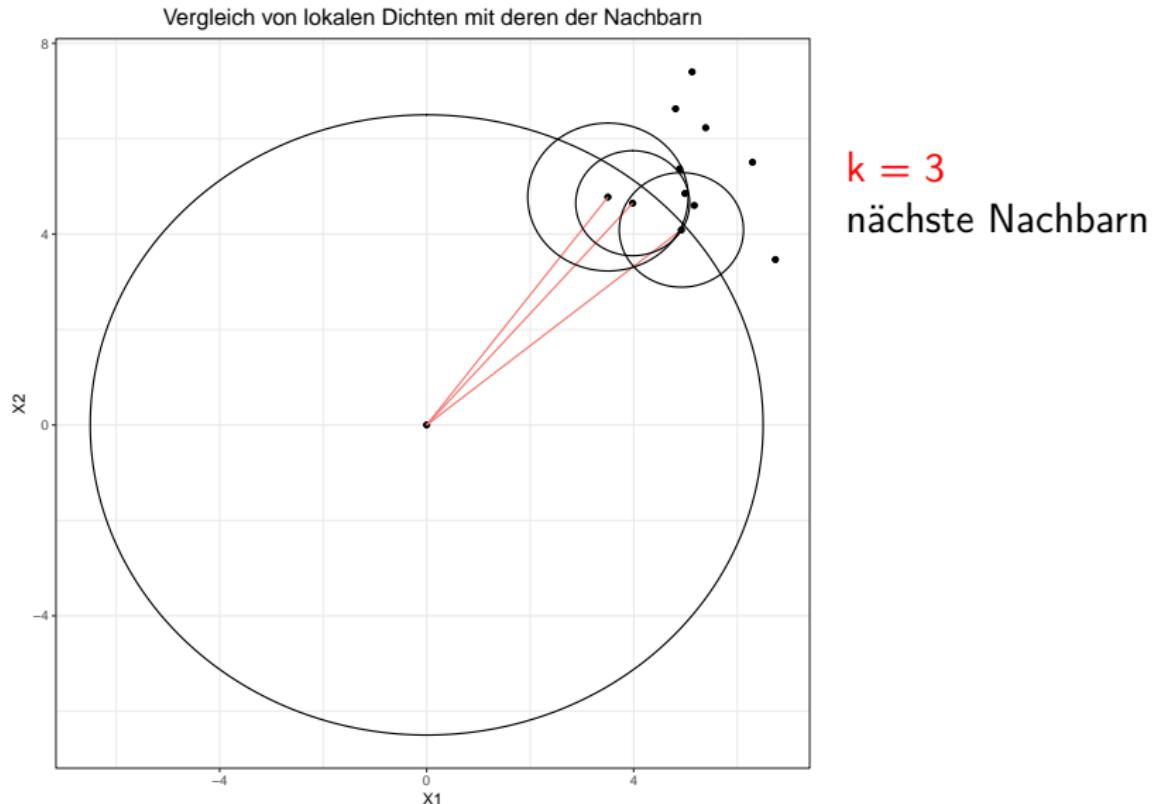


(c) Niedrige Dichte



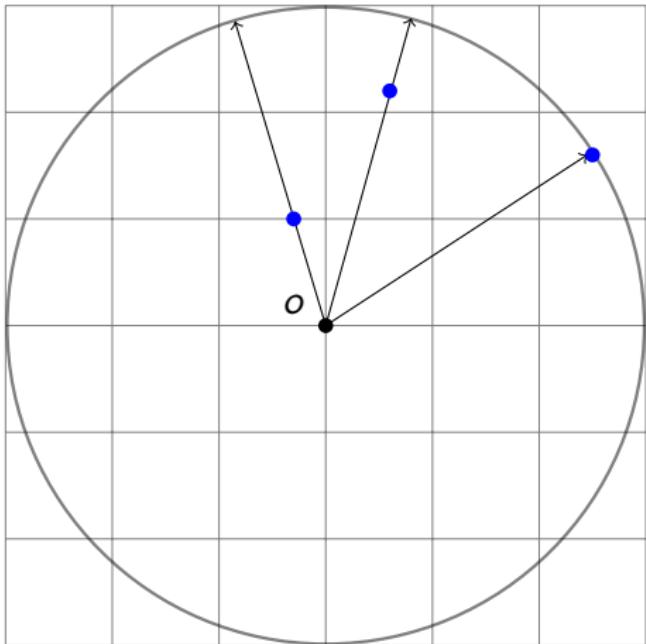
(d) Hohe Dichte

Dichten der nächsten Nachbarn



Konzept der Erreichbarkeitsdistanz

$k = 3$
k-Distanz
k-Dist. Nachbarschaft



$$\text{reach-dist}_k(o, p) = \max\{k\text{-distance}(p), d(o, p)\}$$

Erreichbarkeitsdichte und LOF

- ▶ Zur Bestimmung der LOF werden sogenannte Erreichbarkeitsdichten der Punkte/ Observationen berechnet

$$lrd_k(o) = 1 / \left(\frac{\sum_{p \in N_k(o)} \text{reach-dist}_k(o, p)}{|N_k(o)|} \right)$$

Erreichbarkeitsdichte und LOF

- ▶ Zur Bestimmung der LOF werden sogenannte Erreichbarkeitsdichten der Punkte/ Observationen berechnet

$$Ird_k(o) = 1 / \left(\frac{\sum_{p \in N_k(o)} \text{reach-dist}_k(o, p)}{|N_k(o)|} \right)$$

- ▶ Für den LOF Score werden die Erreichbarkeitsdichten in der Umgebung ins Verhältnis gesetzt

$$\begin{aligned} LOF_k(o) &= \frac{\sum_{p \in N_k(o)} \frac{Ird_k(p)}{Ird_k(o)}}{|N_k(o)|} \\ &= \frac{\frac{1}{Ird_k(o)} \sum_{p \in N_k(o)} Ird_k(p)}{|N_k(o)|} \end{aligned}$$

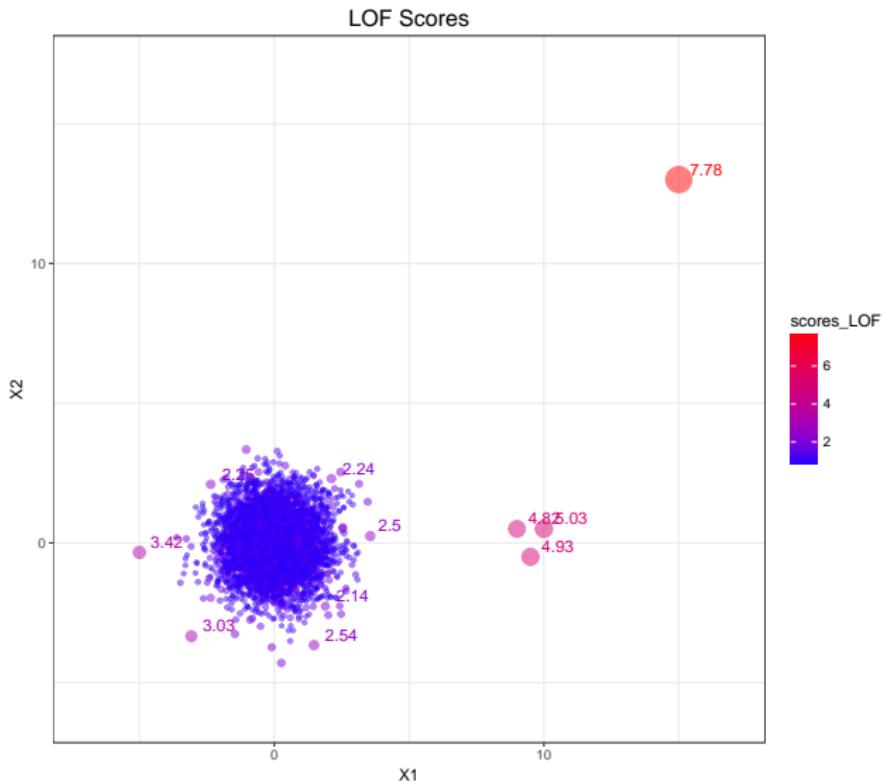
Scores der Local Outlier Factor Methode

- ▶ Je größer die Unterschiede der Dichten sind, desto größer ist der LOF Score

Tabelle: LOF Scores und Interpretation.

LOF Scores	Dichte	Anomalie?
$LOF_k(o) \sim 1$	Vergleichbare Dichte zu Nachbarn	Nein
$LOF_k(o) > 1$	Geringere Dichte als Nachbarn	(Eher) Ja

LOF Scores graphisch



Vor- und Nachteile der LOF Methode

Vorteile:

- ▶ Stabile Methode
- ▶ Lokaler Ansatz kann mit Bereichen unterschiedlicher Dichte umgehen

Vor- und Nachteile der LOF Methode

Vorteile:

- ▶ Stabile Methode
- ▶ Lokaler Ansatz kann mit Bereichen unterschiedlicher Dichte umgehen

Nachteile:

- ▶ Scores schwer zu interpretieren jenseits der 1
- ▶ Kein fester Schwellenwert, ab wann eine Anomalie angenommen wird

Bewertung der Methoden

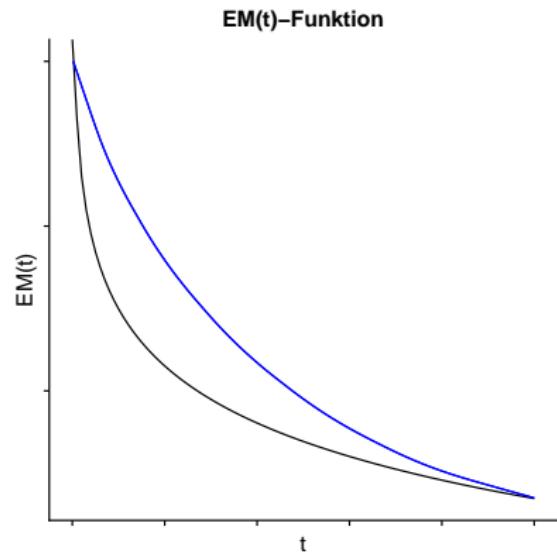
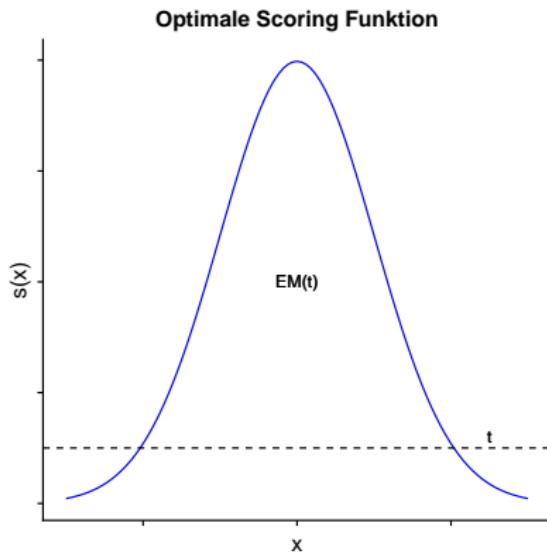
- ▶ Bewertungskriterium für die Performance, das keine gelabelten Daten benötigt
- ▶ Kein Optimieren eines messbaren Fehlers möglich
- ▶ Scoring Funktionen wie iForest und LOF ergeben ein Ranking der Observationen
- ▶ Ansatz für eine Evaluation ist die Scoring Funktion

Excess-Mass (EM) Kriterium

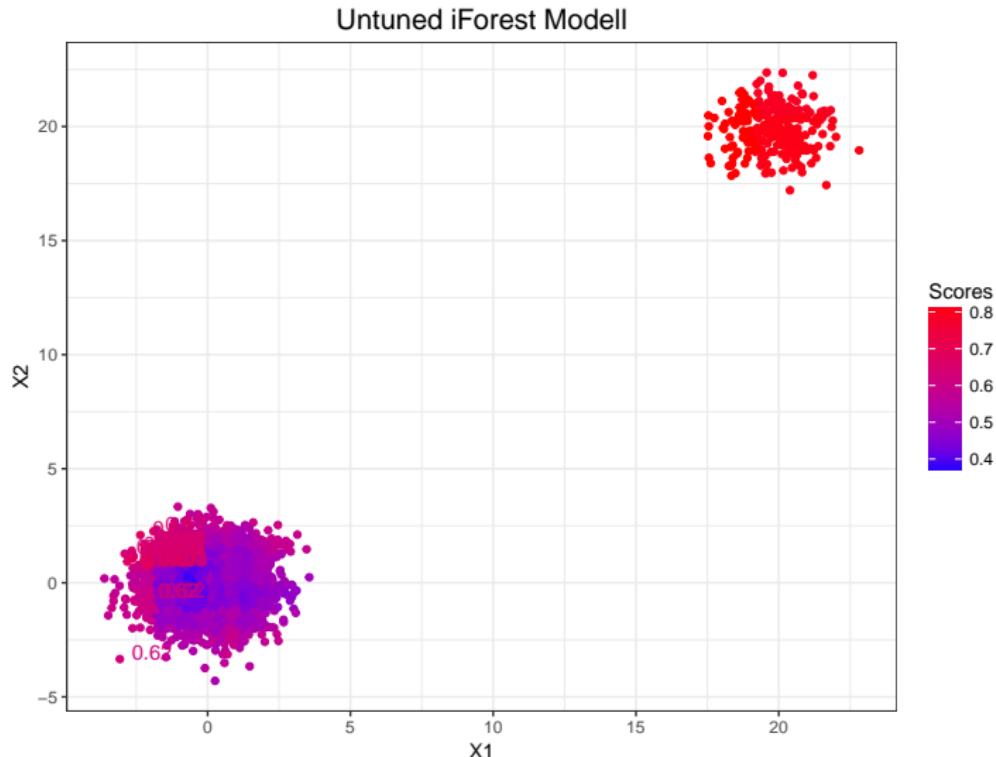
- ▶ Normale Scores sind ähnlich und viele, anomale Scores selten und anders
- ▶ Optimale Scoring Funktion denkbar als wahre Dichte der Daten oder ihre Transformation (streng monoton wachsend)
- ▶ Bewertungskriterium soll Ähnlichkeit der Scoring Funktion zur Dichte messen können
- ▶ Ansatz: Excess-Mass (EM) und Mass-Volume (MV) Kurven
- ▶ Voraussetzung sind stetige Daten

Excess-Mass (EM) Kriterium

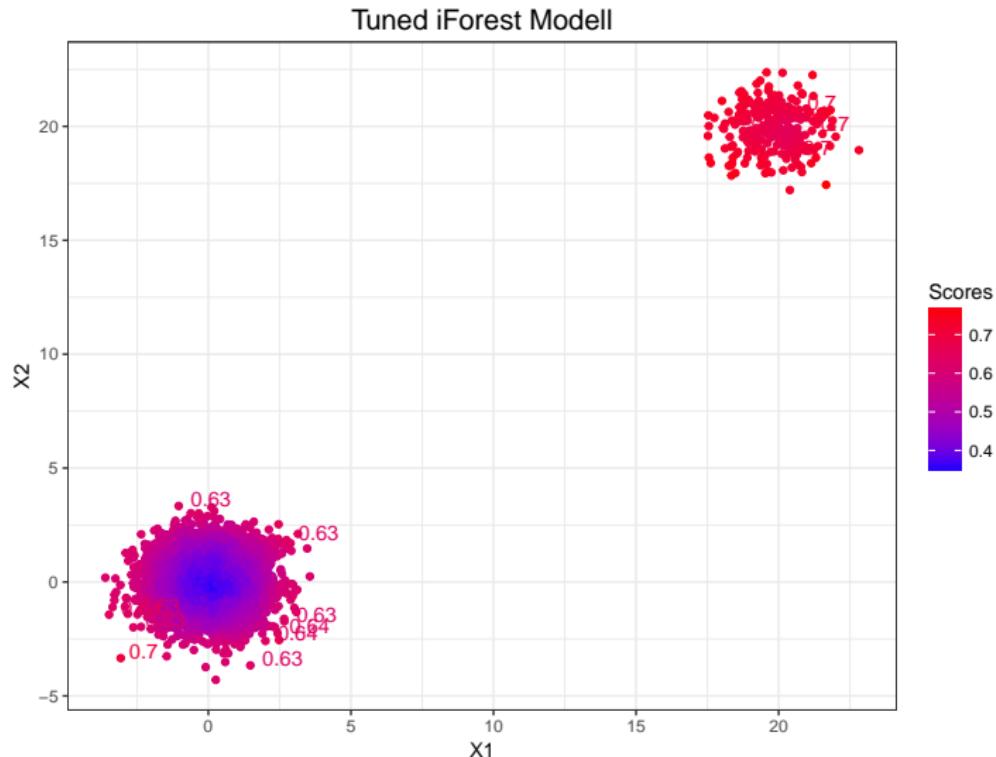
- ▶ Maximierungsproblem auf Basis der Verteilung der Scores der Observationen
- ▶ Höherer Wert $EM(t)$ zeigt relative Güte der Scoring Funktion



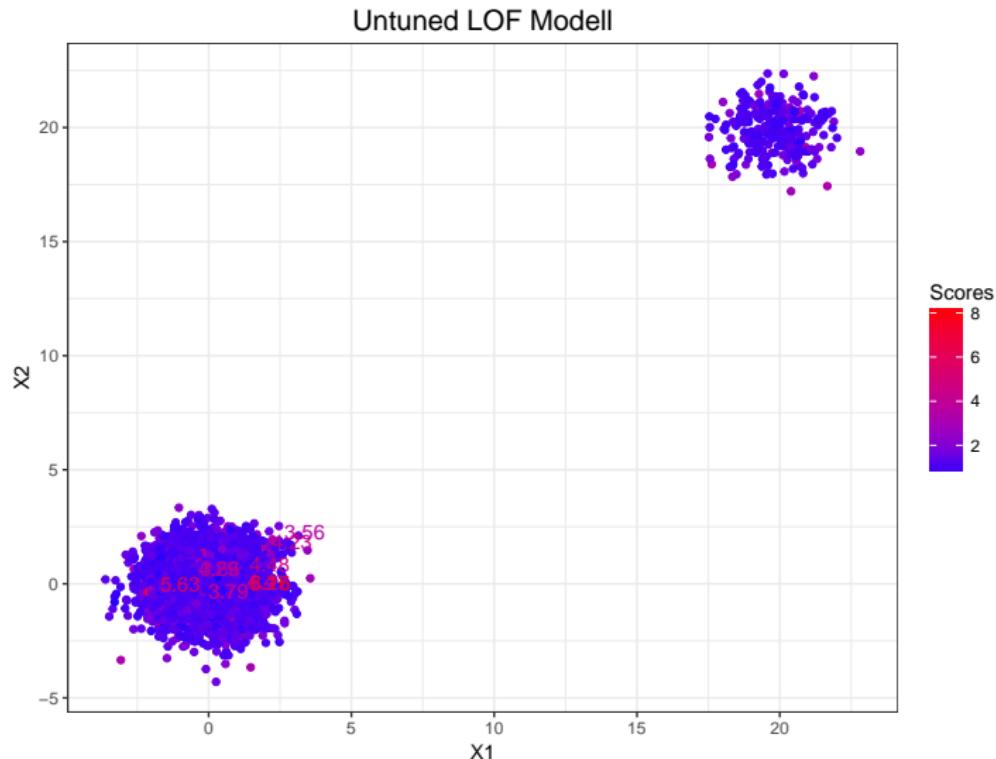
Scores durch einen ungetunten iForest



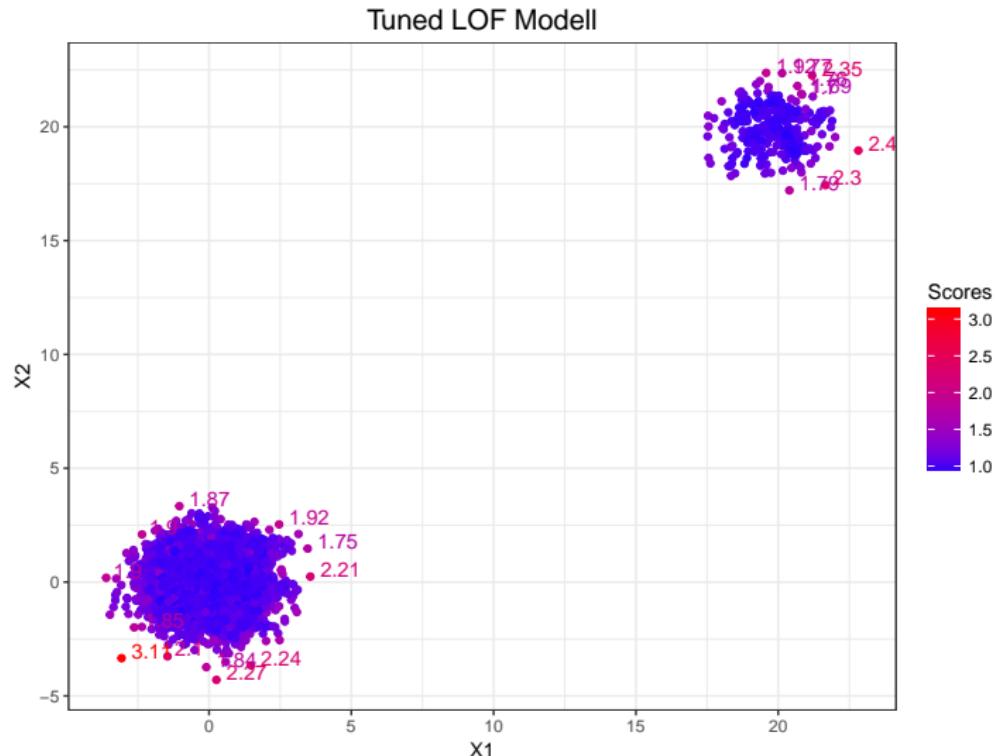
Scores durch einen getunten iForest



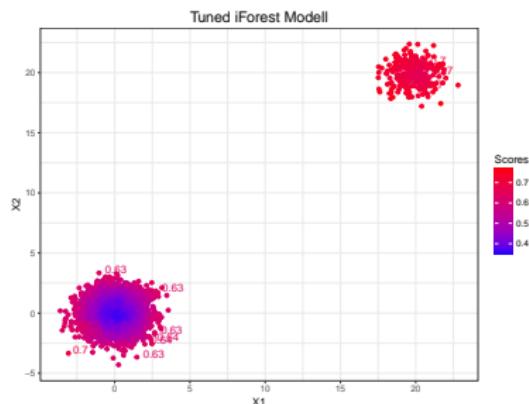
Scores durch die ungetunte LOF Methode



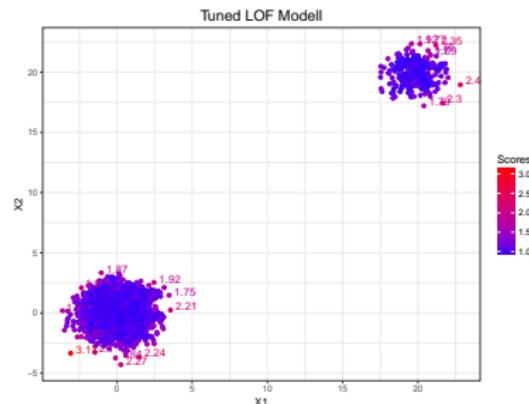
Scores durch die getunte LOF Methode



Vergleich von iForest und LOF Methode



(e) Tuned iForest



(f) Tuned LOF

Vergleich von iForest und LOF Methode

Auf welche Probleme sind wir gestoßen

- ▶ Zeitreihenansatz
- ▶ Kontaminierter Datensatz
- ▶ LOF: Bei identischen Beobachtungen können keine Scores berechnet werden

Gliederung

Einleitung

Projektpartner und Aufgabenstellung
Datengrundlage

Anomaly Detection

Anomaly Detection Methoden
Bewertung von unsupervised Anomaly Detection Methoden
Probleme

Übersicht zu erhaltenen Daten

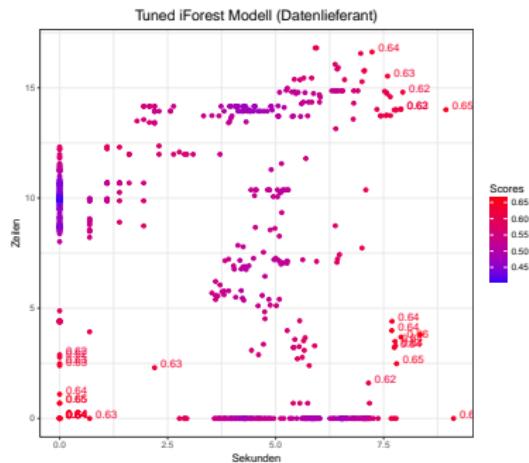
Produkt für Projektpartner und Zusammenfassung

Wiederholung: Datengrundlage

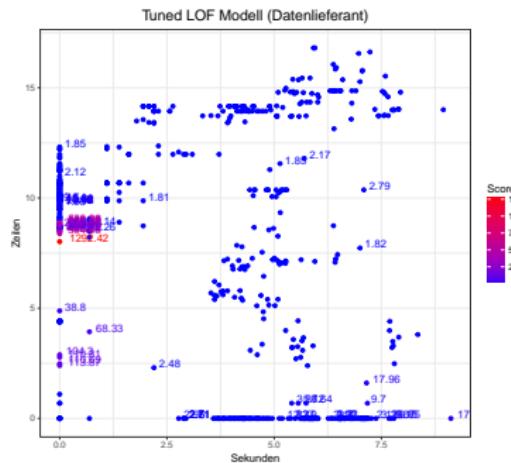
- ▶ Log-Dateien der Ladeprozesse
- ▶ Zeitraum: März/April - November 2017
- ▶ Datensatz mit 9 Variablen
- ▶ Relevant für Anomaly Detection:
 - ▶ „Sekunden“ für Dauer,
 - ▶ „Zeilen“ für Umfang des Ladeprozesses

Startzeit	...	Sekunden	Zeilen	Job	Package	...
...						
...						
<i>9.679 Zeilen</i>						

Vergleich von iForest und LOF Methode



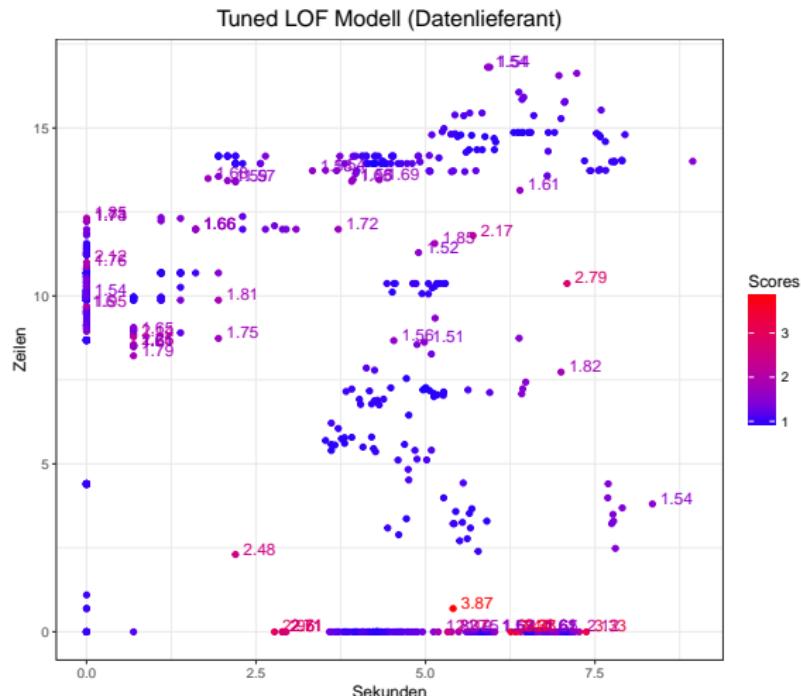
(a) Tuned iForest für Datenlieferant



(b) Tuned LOF für Datenlieferant

Vergleich von iForest und LOF Methode für unsere Daten

LOF Scores für Datenlieferanten ohne extreme Scores



Gliederung

Einleitung

Projektpartner und Aufgabenstellung

Datengrundlage

Anomaly Detection

Anomaly Detection Methoden

Bewertung von unsupervised Anomaly Detection Methoden

Probleme

Übersicht zu erhaltenen Daten

Produkt für Projektpartner und Zusammenfassung

Produkt für Projektpartner

Eine Funktion in R zur Findung von optimalen Parametern für iForest und LOF in Hinsicht auf das EM und/oder MV Kriterium:

```
Anomaly_Detection(data, MV = FALSE, EM = TRUE, Forest = TRUE,  
                    Rand.search = TRUE, min_ntree = 20, max_ntree,  
                    SSize = nrow(data), LOF = TRUE, min_nn = 2, max_nn,  
                    mc.cores = 1, n_generated = 1000, t_max = 0.9,  
                    alpha_min = 0.6, alpha_max = 0.999, ...)
```

Zusammenfassung

- ▶ Anomalien sind anders und selten
- ▶ Verwendete Methoden iForest und LOF erstellen ein Ranking der Observationen
- ▶ EM Kriterium bewertet Güte von Scoring Funktionen
- ▶ Tuning auf Basis des EM Kriteriums verbessert die Performance
- ▶ R-Funktion liefert optimierte Parameter für Anomaly Detection Modelle
- ▶ Projektpartner kann abgeschlossene Ladeprozesse überprüfen

Ausblick

- ▶ „Gebrauchsanweisung“ für Projektpartner
- ▶ Offene Punkte und zusätzliche Funktionen
- ▶ Weitere Ansätze

Ausgewählte Literatur

-  Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander (2000)
LOF: Identifying Density-Based Local Outliers
<http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf>
-  Nicolas Goix, Anne Sabourin & Stephan Clemenccon (2015)
On Anomaly Ranking and Excess-Mass Curves
<http://proceedings.mlr.press/v38/goix15.pdf>
-  Nicolas Goix (2016)
How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?
<https://arxiv.org/pdf/1607.01152.pdf>
-  Fei Tony Liu, Kai Ming Ting & Zhi-Hua Zhou (2013)
Isolation Forest
<https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf>

Anhang

EM-Kurve

$$EM_s(t) = \sup_{u \geq 0} \mathbb{P}(s(\mathbf{X}) \geq u) - t * Leb(s \geq u)$$