Meilenstein 2

Verlustprävention an Selbstbedienungskassen im Einzelhandel

Durchgeführt durch die
Retail Data Mining GmbH

RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

Themen für heute:

1. **Vorbemerkung zur Datenauswertung**
2. **Grundlegende Datenanalyse**
3. **Datentransformation**
4. **Explorative Datenanalyse**
5. **Fazit und Ausblick**

RDM
Retail Data Mining GmbH

1. Vorbemerkung

Datenannahmen & potenzielle Verzerrungen

- Annahme: **KassenSichV-konforme** Daten (manipulationssicher)

- Mögliche Verzerrung: **Keine negativen Schadensfälle im Datensatz**
  - z. B. vergessene Ware → theoretischer Überzahlung
  - Relevanz für Nettoverlust- und Risikobetrachtungen

- Keine ergänzenden **Auditberichte** oder **technische Dokumentationen** verfügbar

RDM
Retail Data Mining GmbH

Fokus des zweiten Meilensteins

- Fokus: Datenaufbereitung, Management & EDA (Explorative Datenanalyse)

- Vorbereitung für nachfolgende Modellierungsphasen: „**Exploration before prediction**“ – solide Basis für belastbare Modelle

- Teil des **iterativen Vorgehens** nach DASC-PM
  - Analysen & Modelle werden bei Bedarf angepasst
  - Neue Erkenntnisse oder zusätzliche Daten → Re-Validierung möglich

RDM
Retail Data Mining GmbH

2. Grundlagen

Repräsentativität

Vergleich klassifizierter („gelabelter“, d.h. „FRAUD“ bzw. „NORMAL“) Daten mit dem restlichen Datensatz:

Numerische Spalten (t-Tests):

	Spalte	p-Wert	Mittelwert (labeled)	Mittelwert (unlabeled)	Std-Abw (labeled)	Std-Abw (unlabeled)
3	transaction_duration	0.185389	77.807475	77.541994	73.202614	72.895636
1	n_lines	0.355874	10.603607	10.575406	11.155176	11.101239
2	customer_feedback	0.671868	9.326005	9.318636	1.699571	1.715356
0	total_amount	0.750073	98.509750	98.413698	110.079582	109.943709

(Ausschnitt für numerische Merkmale auf Basis eines t-Tests)

Fazit: **gelabelte Daten sind repräsentativ** für den gesamten Trainingsdatensatz

Achtung: Unterschiede zwischen Trainings- und Testdaten!

RDM
Retail Data Mining GmbH

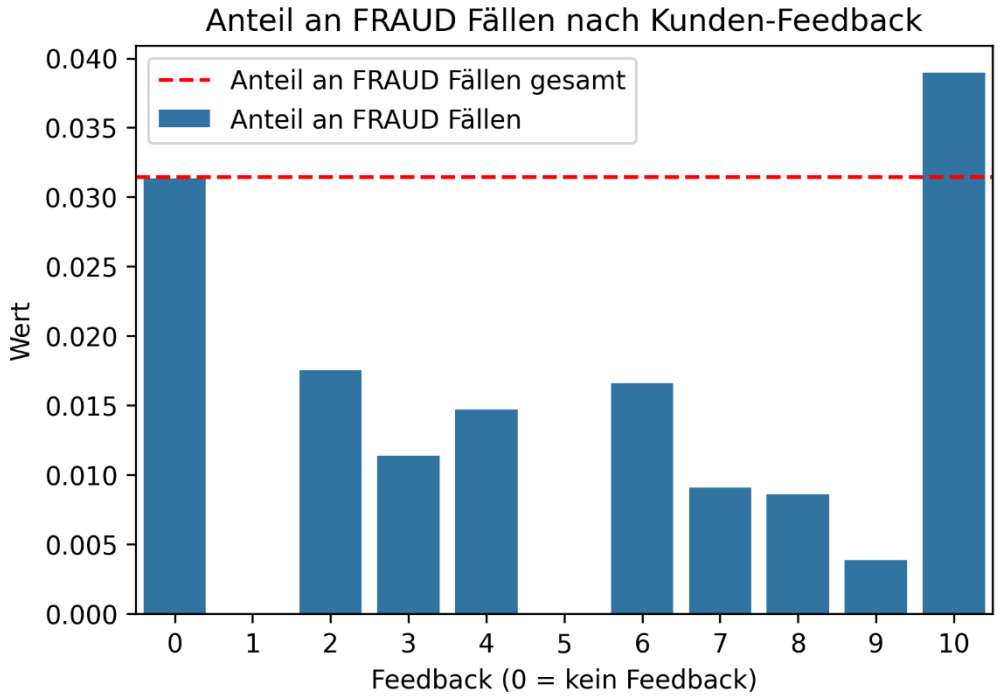
Plausibilität

- Daten im Wesentlichen konsistent, aber:
  - **Komplexe Stornothematik** → konnte in Meilenstein 2 nicht abschließend geklärt werden, muss in Meilenstein 3 erneut aufgenommen werden
  - Durch statische Regeln lassen sich viele als „FRAUD“ deklarierte Transaktionen **sehr sicher vorhersagen**

- Berücksichtigung bei **späterer Modellbildung**

RDM
Retail Data Mining GmbH

Auffälligkeit – Kundenfeedback

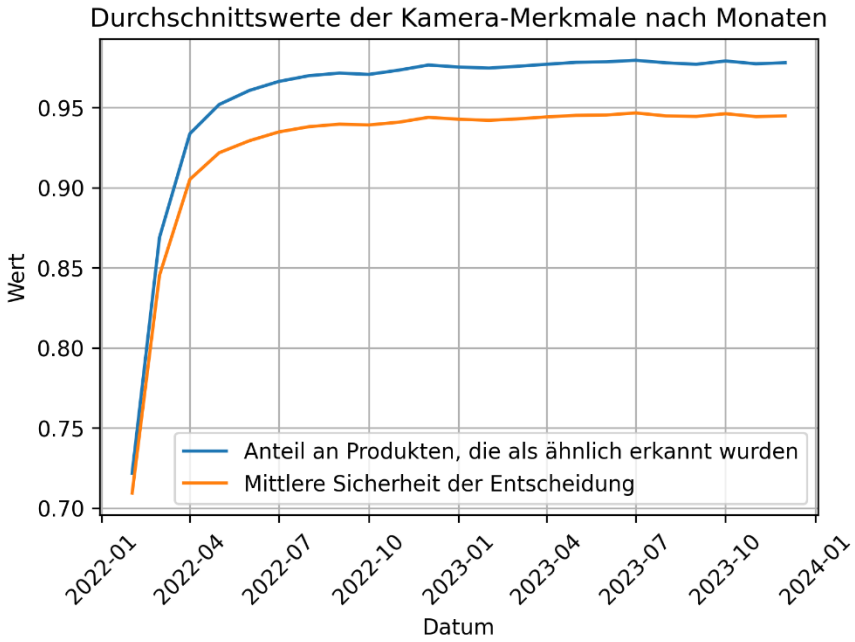


Wenige Werte bei Kundenfeedback und bei vorhandenen Werten extreme Ausprägung (bei Fraud mehrheitlich volle Punktzahl)

RDM
Retail Data Mining GmbH

Lernkurve – Kamerasystem

- **Kamerasystem anfangs nicht ausgelernt**
- Spätere Daten deutlich brauchbarer
- Zu beachten bei zukünftiger Einführung eines neuen Kamerasystems oder bei einer neuen Filiale



RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

3. Datenmanagement

Transformation der Daten

- 4 Datentabellen Tabellen in **eine einzige Datentabelle** überführt
- Relevante Transaktions- und Artikeldaten extrahiert bzw. berechnet
- Formatbereinigung und Überführung in analysierbare Tabellenstruktur
- **Pro Transaktion eine Zeile** erzeugt
- Artikelpositionen je Transaktion **zu Merkmalen aggregiert**

RDM
Retail Data Mining GmbH

Aggregation der Daten

- **Positionsdaten zu Merkmalen aggregiert** (z.B. enthält Snacks, durchschnittliche Scanzeit pro Artikel etc.)

- Sowohl **kategoriale Merkmale** als auch **numerische**:
  - Tritt eine Kategorie in der Transaktion auf? Ja / nein
  - Wie viele Fälle? Anzahl

- **Transformation der Produktkategorien:**
  - Ist eine Produktkategorie vorhanden oder nicht (Getränke, Snacks, usw.)

- Minimum/Maximum/Mittelwert (Preis, Popularität, Zeit zwischen Scans)

RDM
Retail Data Mining GmbH

Umgang mit unvollständigen Daten (1)

- Feedback: nur in **7,6 % der Fälle vorhanden**
  - Transformation zu kategorialen Ausprägungen (sehr gut, gut, mittel, schlecht, überhaupt vorhanden)

- 11.479 Fälle mit fehlenden Werten für mittlere und maximale Zeit zwischen Scans
  - Ursache: Nur ein Scan vorhanden
  - Ersetzt durch Mittelwert

RDM
Retail Data Mining GmbH

Umgang mit unvollständigen Daten

- 114 Fälle mit **fehlenden Werten des Kamerasystems**
  - Ersetzt durch den Modus

- Ein Fall mit mehreren fehlenden Spaltenwerten aufgrund fehlender Produkt-ID → entfernt

- Da wir nur die klassifizierten Daten betrachten → **keine Veränderung der nicht-klassifizierten Daten**

RDM
Retail Data Mining GmbH

4. Explorative Analyse

Übersicht

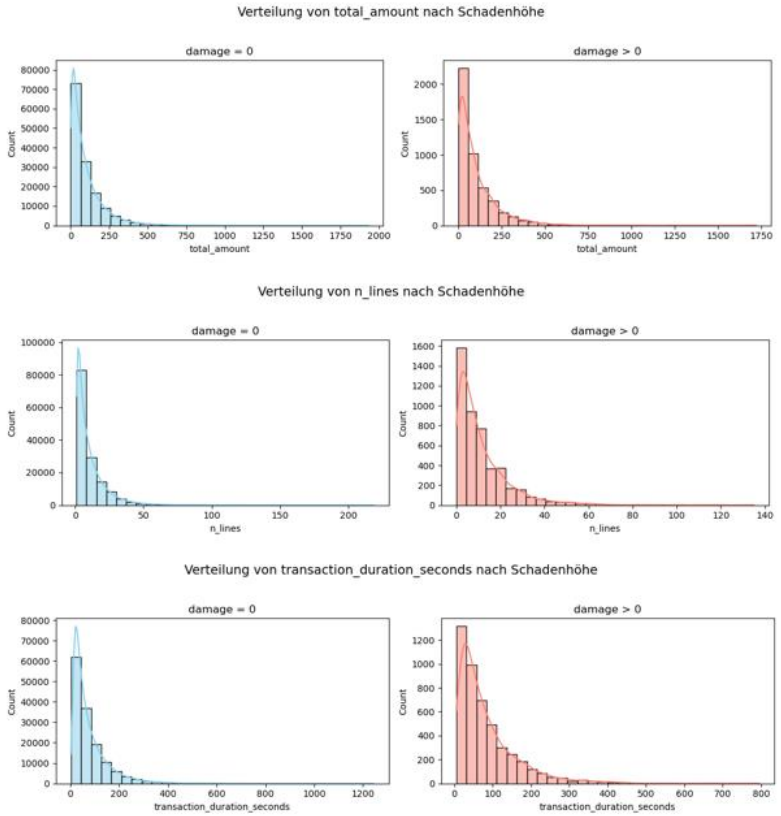
- 4 Schritte in der explorativen Datenanalyse:
  - **Verteilungsanalyse** und Ausreißer **numerischer** Attribute
  - Analyse **kategorialer Attribute**
  - **Nichtlineare Zusammenhänge** zwischen Attributen und Schadenshöhe
  - **Regressionsmodellierung**

RDM
Retail Data Mining GmbH

Numerische Merkmale von FRAUD (1)

- Transaktionen mit Schaden (damage > 0):
  - **höhere Warenkorbsummen**
  - **mehr** gekaufte **Artikel** (n_lines)
  - **längere Transaktionsdauer**

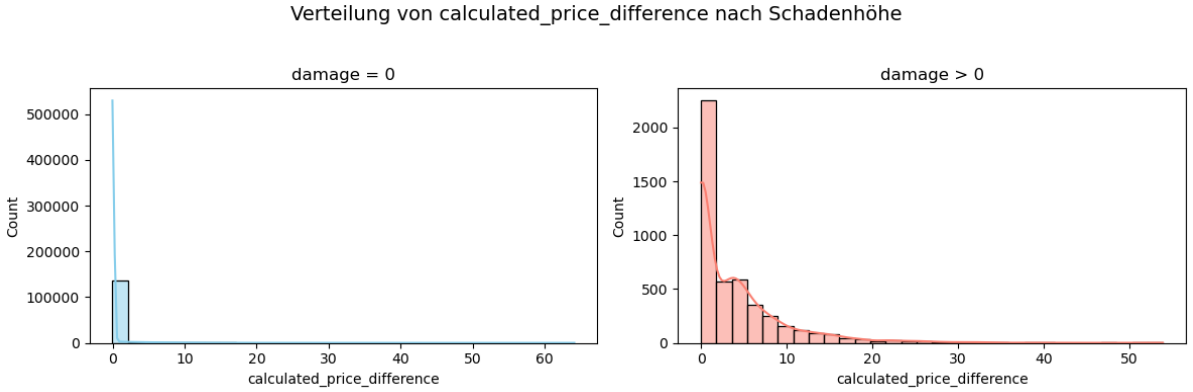
- Merkmale sind **stark korreliert**
- **Interpretation:**
  - Mit wachsendem Warenkorb steigt die Komplexität
  - Fehler wie falsches Scannen oder vergessene Artikel werden wahrscheinlicher



RDM
Retail Data Mining GmbH

Numerische Merkmale von FRAUD (2)

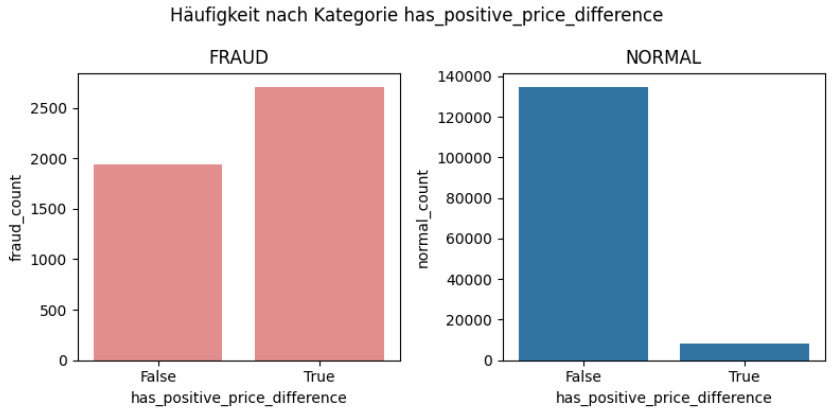
- Transaktionen mit Schaden (damage > 0):
  - deutlich höhere *calculated_price_difference (*Differenz zwischen Summe der Einzelpreise und Kassensumme)
  - *calculated_price_difference* als potenziell **starker Prädiktor** für Verluste



RDM
Retail Data Mining GmbH

Numerische Merkmale von FRAUD (3): Bezahlter Preis ≠ Nominalpreis

- **Nominalpreis** einer **Position**: Menge bzw. Gewicht multipliziert mit dem Nominalpreis des Artikels gemäss Produkttabelle
- **Nominalpreis** einer **Transaktion**: Summe der Nominalpreise aller nicht-stornierten Artikel
- **Häufige Abweichungen**
- Zwei definierte Merkmale:
  - **Differenz vorhanden** (ja/nein)
  - **Absolute Höhe** der Differenz

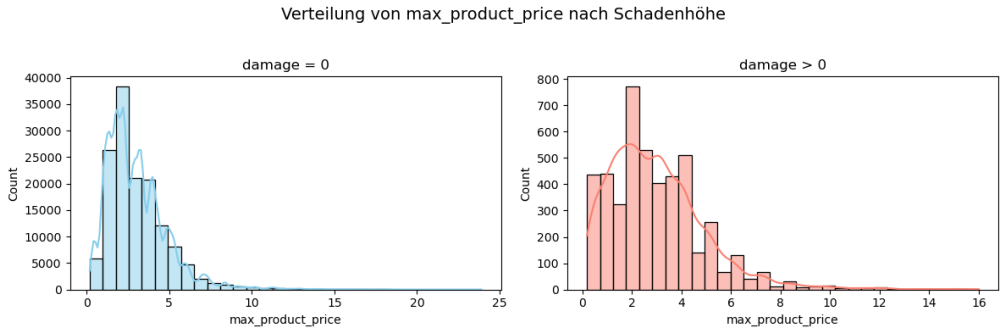


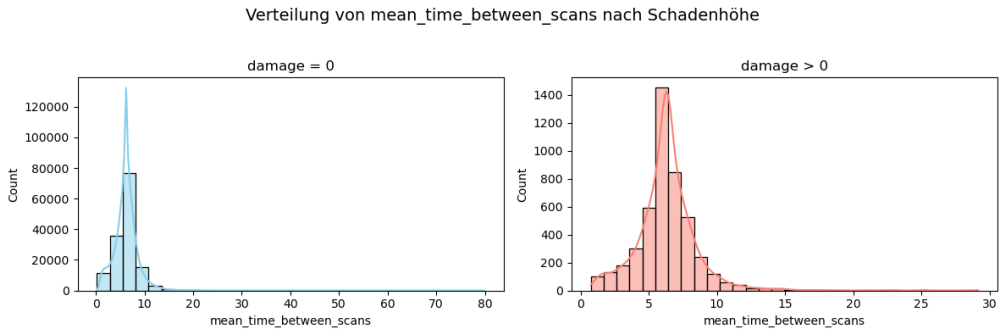
RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

Numerische Merkmale von FRAUD (4)

- Transaktionen mit Schaden (damage > 0):
  - enthalten häufiger **hochpreisige Einzelartikel**
  - **breitere Streuung** bei der mittleren Zeit zwischen Scans





RDM
Retail Data Mining GmbH

Numerische Merkmale: Extremwerte

- Für alle numerischen Features wurde der **Z-Score** berechnet
- Nutzen: Identifikation systematisch **auffälliger Attribute**
- Interpretation: Extremwerte nicht als Rauschen, sondern als **potenziell erklärungsstark** anzusehen

feature	outliers_abs_zscore>3
calculated_price_difference	3273
popularity_max	3193
total_amount	2962
ransaction_duration_seconds	2947
n_lines	2906
max_time_between_scans	2204
time_from_last_scan_to_end	2167
damage	2111
max_product_price	2073
mean_time_between_scans	1386
time_to_first_scan	949
popularity_min	161
days_since_sco_introduction	0

RDM
Retail Data Mining GmbH

Numerische Merkmale: Signifikanz

- t-Test als Entscheidungskriterium, welche Prädiktoren signifikant sind
- Zusätzlich Analyse, wie viel mit dem Prädiktor erklärt werden kann (Relevanz)

feature	significance	relevance
payment_medium	sehr signifikant	sehr relevant
hour	sehr signifikant	weniger relevant
[redacted]		
has_voided	sehr signifikant	weniger relevant
n_voided	sehr signifikant	weniger relevant
has_camera_detected_wrong_produc	sehr signifikant	weniger relevant
calculated_price_difference	sehr signifikant	sehr relevant
has_positive_price_difference	sehr signifikant	weniger relevant
has_snacks	sehr signifikant	weniger relevant

RDM
Retail Data Mining GmbH

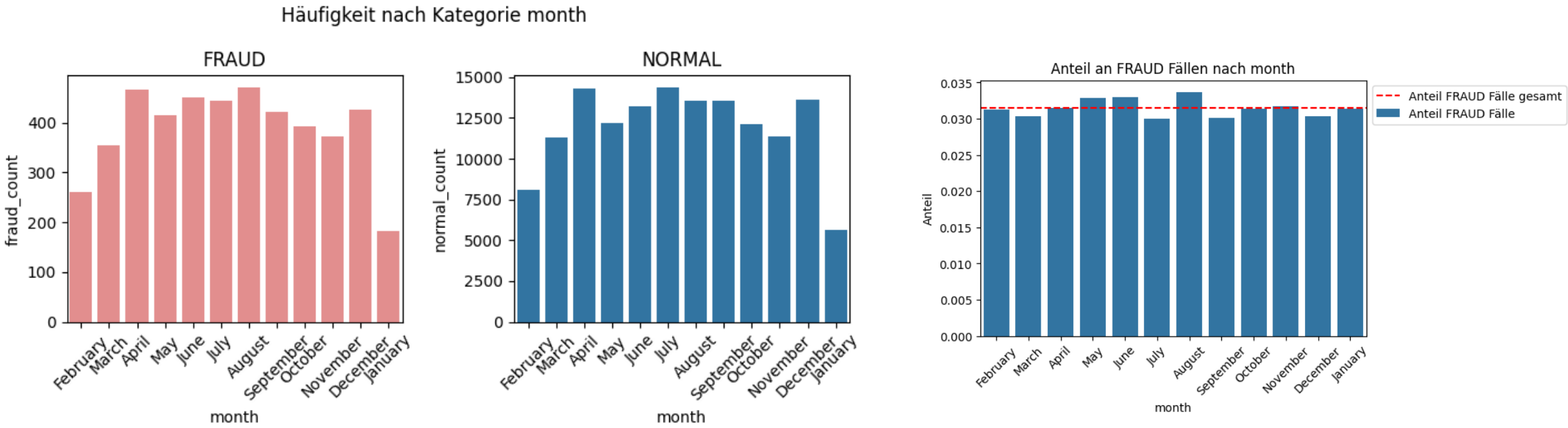
Kategoriale Merkmale von Fraud (1)

- Im Folgenden einige graphische Gegenüberstellungen von FRAUD / NORMAL anhand kategorialer Variablen

- Insbesondere bestimmte Produktkategorien kommen hier besonders häufig vor, ebenso:
  - Wurde mehrheitlich bar bezahlt
  - Hat das Kamerasystem Auffälligkeiten bemerkt

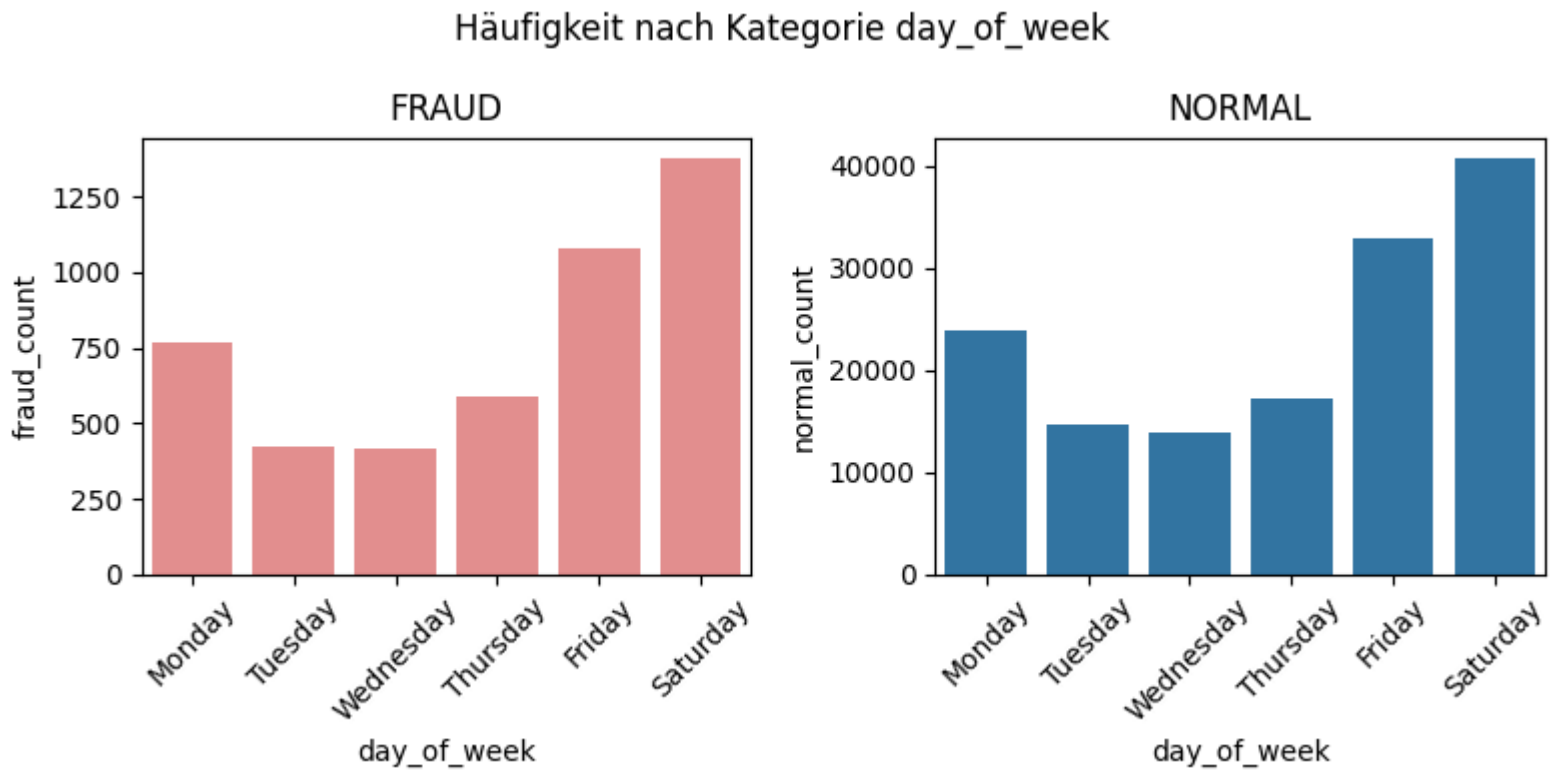
RDM
Retail Data Mining GmbH

Kategoriale Merkmale: Monat



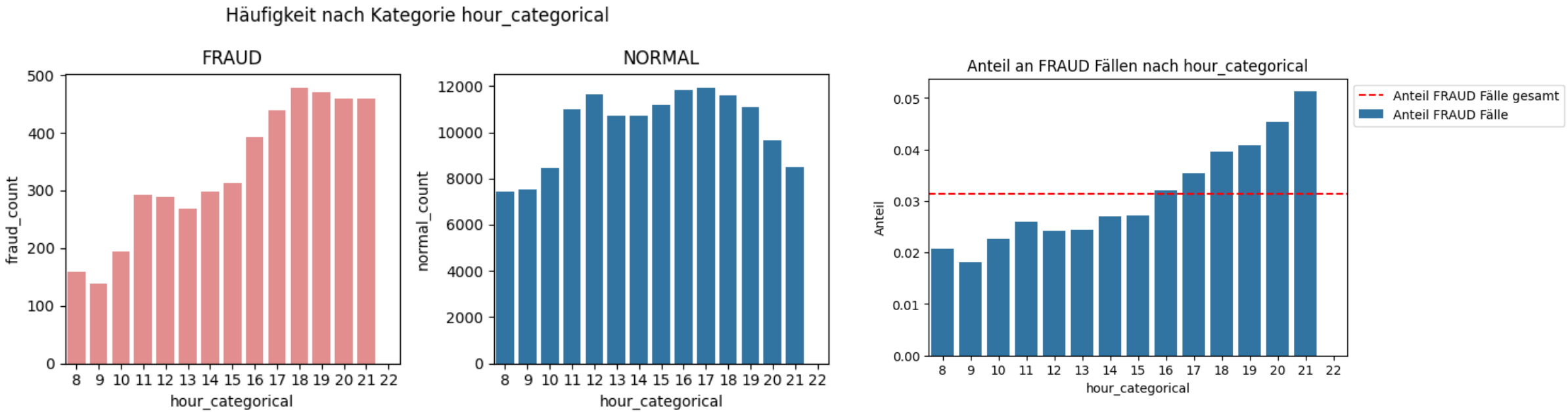
RDM
Retail Data Mining GmbH

Kategoriale Merkmale: Wochentag



RDM
Retail Data Mining GmbH

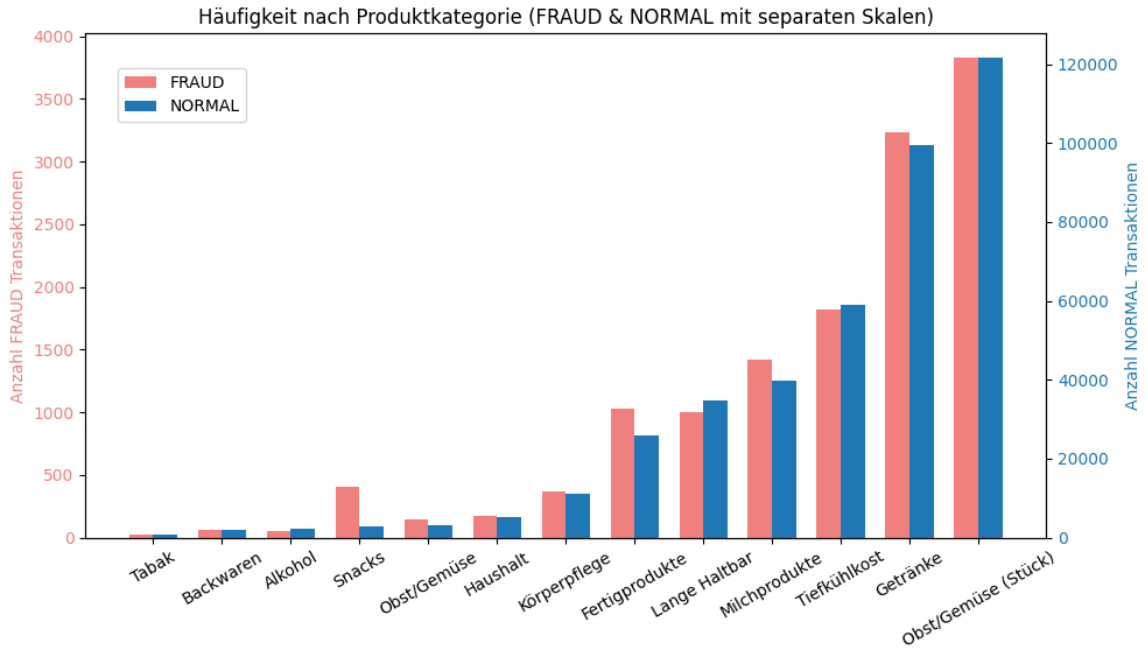
Kategoriale Merkmale: Tageszeit

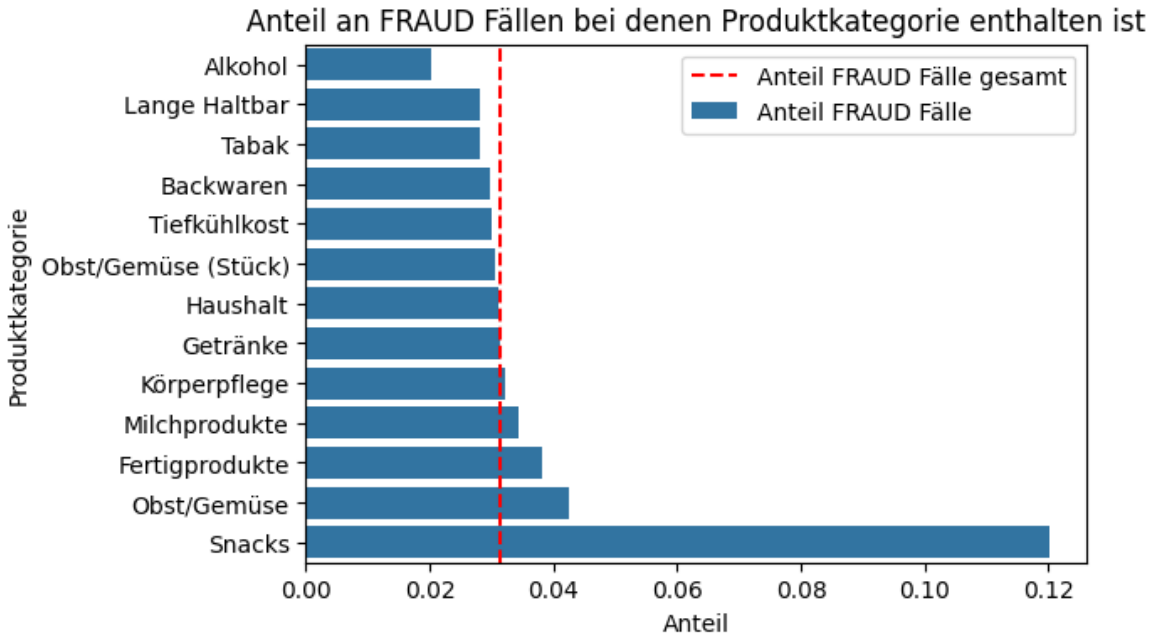


RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

Kategoriale Merkmale: Produktkategorie

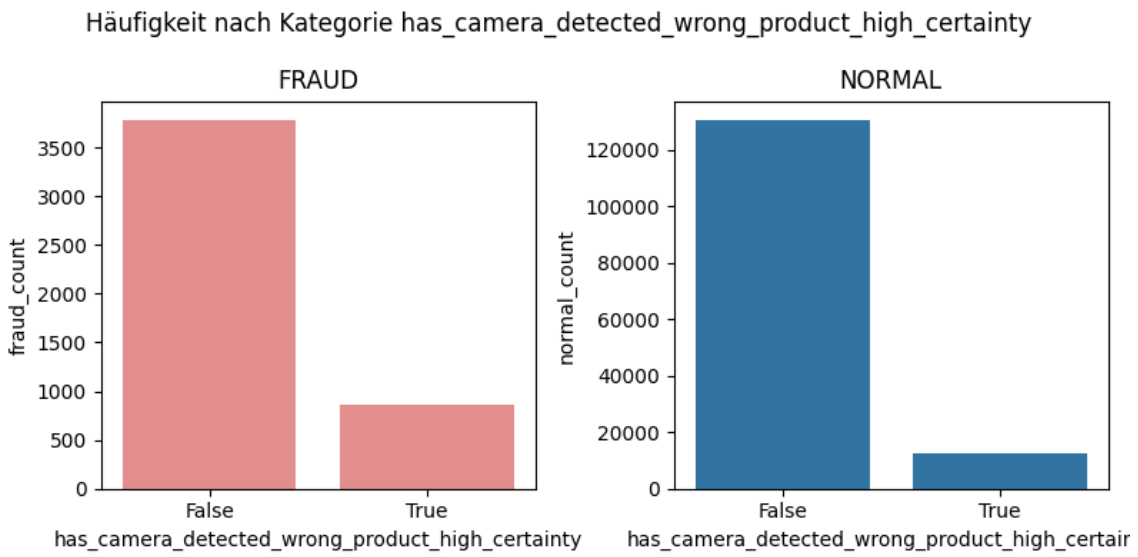


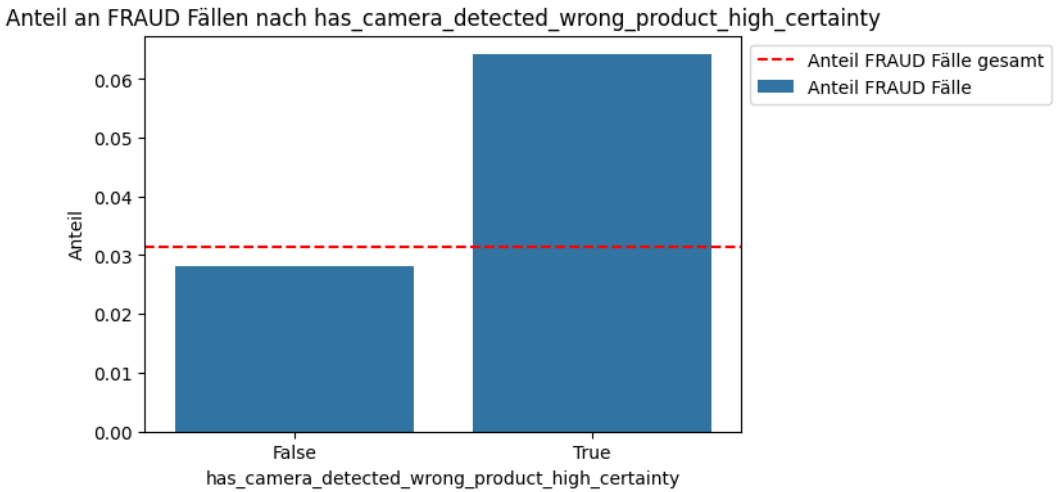


RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

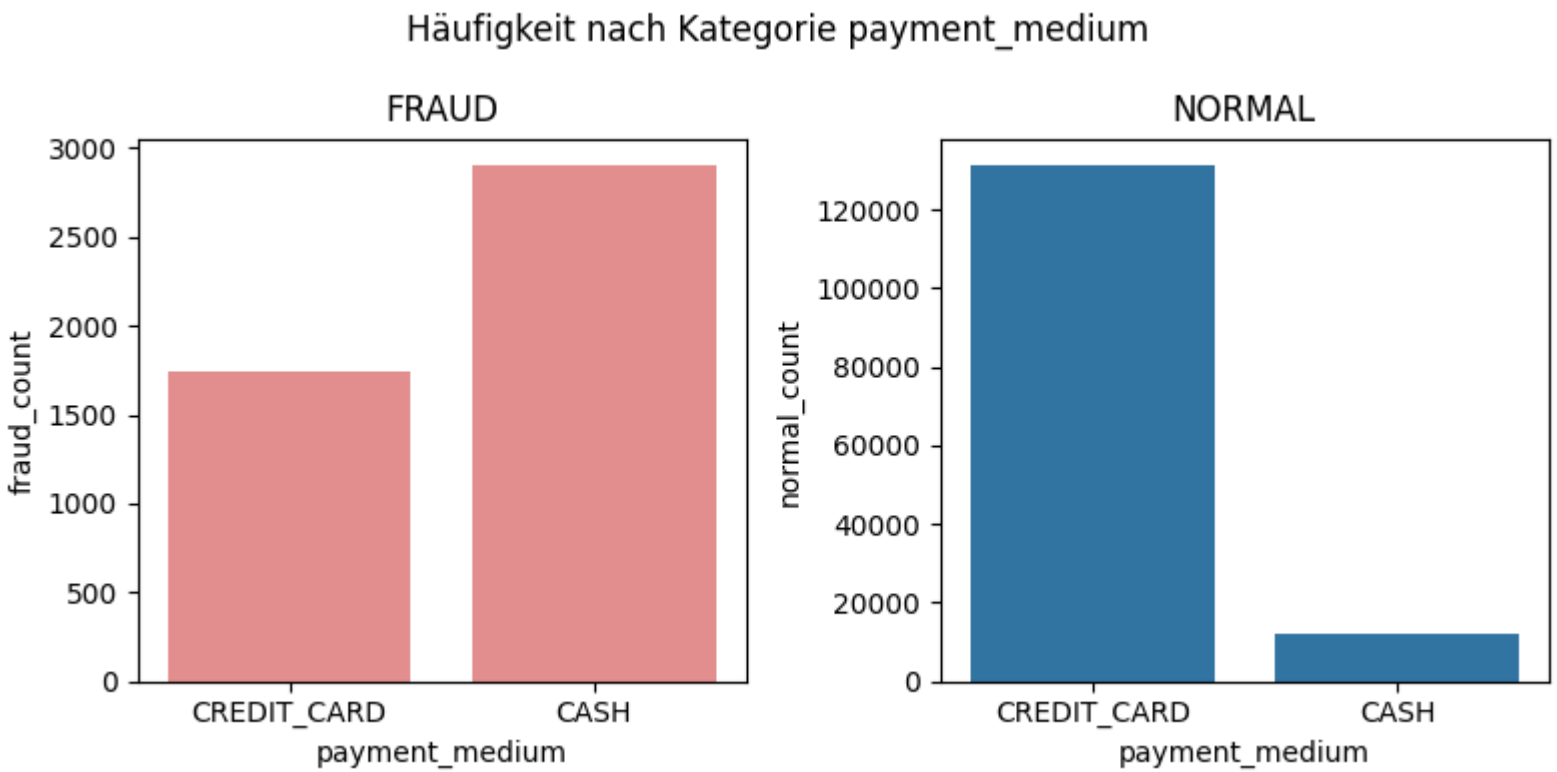
Kategoriale Merkmale: Kamerasystem





RDM
Retail Data Mining GmbH

Kategoriale Merkmale: Zahlungsmittel



RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

Kategoriale Merkmale: Signifikanz

- Chi^2-Test als Entscheidungskriterium, welche Prädiktoren signifikant sind
- Zusätzlich Analyse, wie viel mit dem Prädiktor erklärt werden kann (Relevanz)

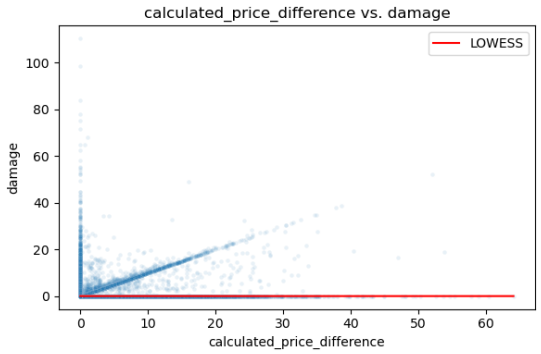
feature	significance	relevance
payment_medium	sehr signifikant	sehr relevant
calculated_price_difference	sehr signifikant	weniger relevant
has_positive_price_difference	sehr signifikant	sehr relevant

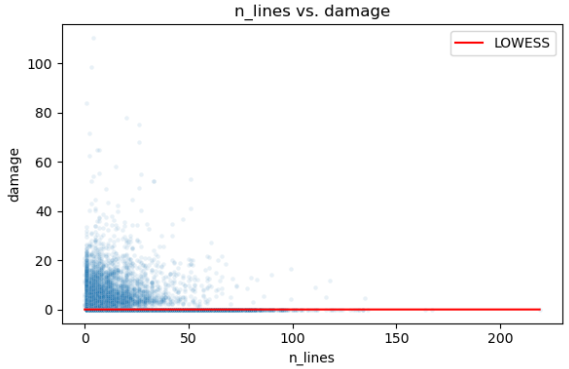
RDM
Retail Data Mining GmbH

Nichtlineare Zusammenhänge

- Zur Analyse nichtlinearer Zusammenhänge zwischen numerischen Attributen und Schadenshöhe zwei Ansätze:
  - **LOWESS-Glättung** zur visuellen Trendbewertung
  - **Spearman & Pearson-Korrelation** zur quantitativen Bewertung

- Ergebnisse: Die meisten Merkmale zeigen keine klare nichtlineare Beziehung. Lediglich zwei Merkmale zeigen komplexere Beziehung zur Schadenshöhe.





RDM
Retail Data Mining GmbH

Regressionsanalyse: Multivariate Analyse

- **Multivariate Modellbildung** mit Reduktion (schrittweise Entfernen nicht relevanter Attribute)
- Getrennte Betrachtung für Zielgrößen:
  - Logistische Regression: FRAUD / NORMAL
  - Klassische Regression: Schadenshöhe

- Aufteilung der Daten in eine Trainingsmenge (80%) und eine Validierungsmenge (20%). Bewertung anhand der Performance auf beiden Mengen.

RDM
Retail Data Mining GmbH

Regressionsanalyse: Auswertung

- Prognosegüte bei Klassifikation ist **verzerrt** durch die vielen Nicht-Schadensfälle; **bei ausgewogenem Datensatz bessere Performance**

- Geringe Vorhersagbarkeit der Schadenshöhe
  - **Breite Streuung** der Schadensbeträge
  - Großer Anteil an Null-Schäden → Verteilung verzerrt

- Komplexere Modelle mit Interaktionen:
  - **Verbesserung auf Trainingsdaten**, aber
  - **Kein Zugewinn auf Testdaten** → Überanpassung

```
Label-Modell:
Accuracy Test:  0.974
Accuracy Train: 0.974

Confusion Matrix Test:
Predicted    0.0  1.0
Actual
0.0        28646   43
1.0          712  204

Confusion Matrix Train:
Predicted     0.0  1.0
Actual
0.0        114498  182
1.0          2860  879
```

```
Damage-Modell:
R² Test:  0.137
R² Train: 0.136

RMSE Test:  1.754
RMSE Train: 1.721
```

RDM
Retail Data Mining GmbH

5. Ausblick

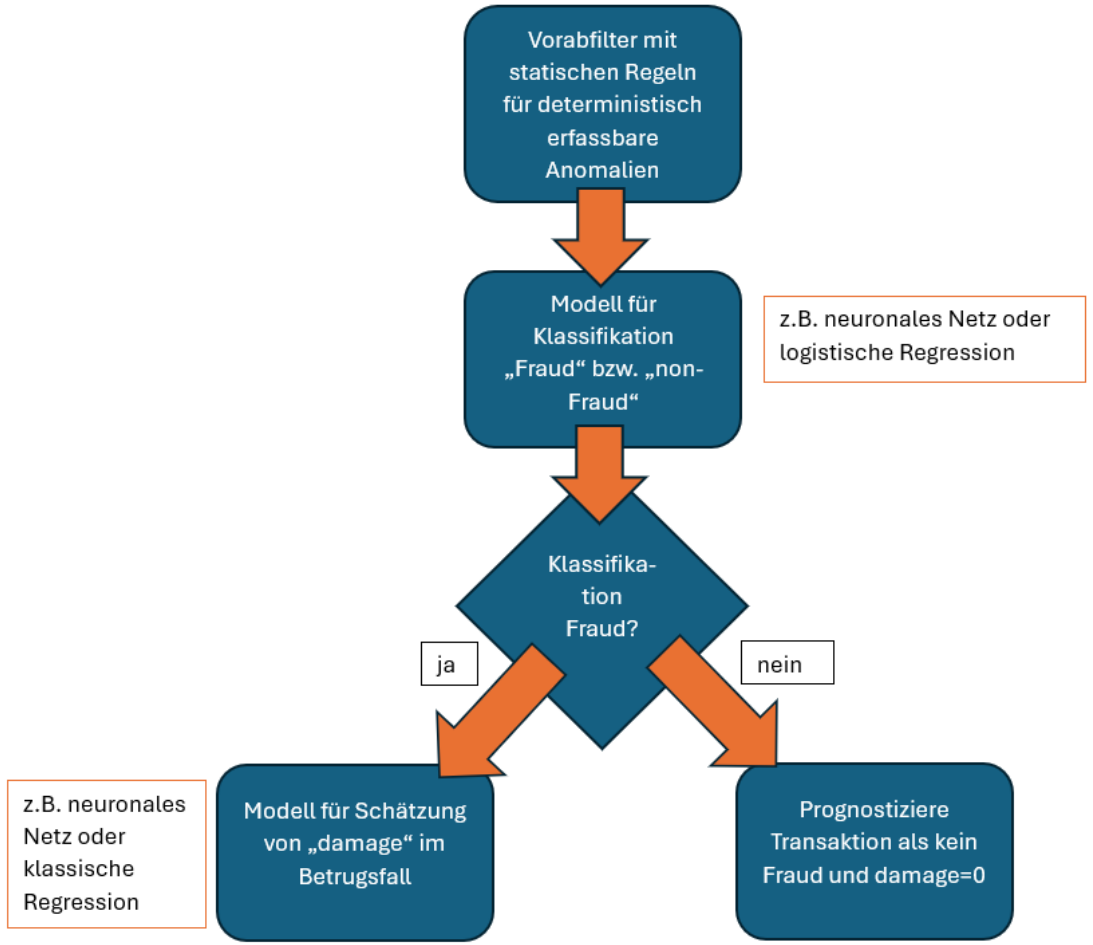
Fazit des zweiten Meilensteins

- Daten sind **plausibel und konsistent** (Stornothematik noch zu klären)
- **Relevante Merkmale** wurden extrahiert und statistische analysiert
- Daten eignen sich für **weiteren Modellaufbau**
- **Komplexere Verfahren notwendig**, um durchgehend gute Prognosegüte sowohl bei der Klassifikation als auch der Schadensvorhersage gut abzuschneiden

RDM
Retail Data Mining GmbH

Nächste Schritte

- **Dreistufiges Modell** auf Grundlage der aktuellen Datenerkenntnisse:
  - 1. Statische Anwendung gewisser **Erkennungsregeln**
  - 2. **Klassifikationsalgorithmus** zur Erkennung von fehlerhaften Transaktionen
  - 3. **Modell für Schätzung der Schadenshöhe** im Falle fehlerhafter Transaktionen (ansonsten prognostiziere Schaden=0)
  - Einbau der **Bewertungsfunktion** in Regeln für manuelle Kontrollen



RDM
Retail Data Mining GmbH

Vielen Dank für Ihre Aufmerksamkeit!

Fragen & Anregungen?

RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik