

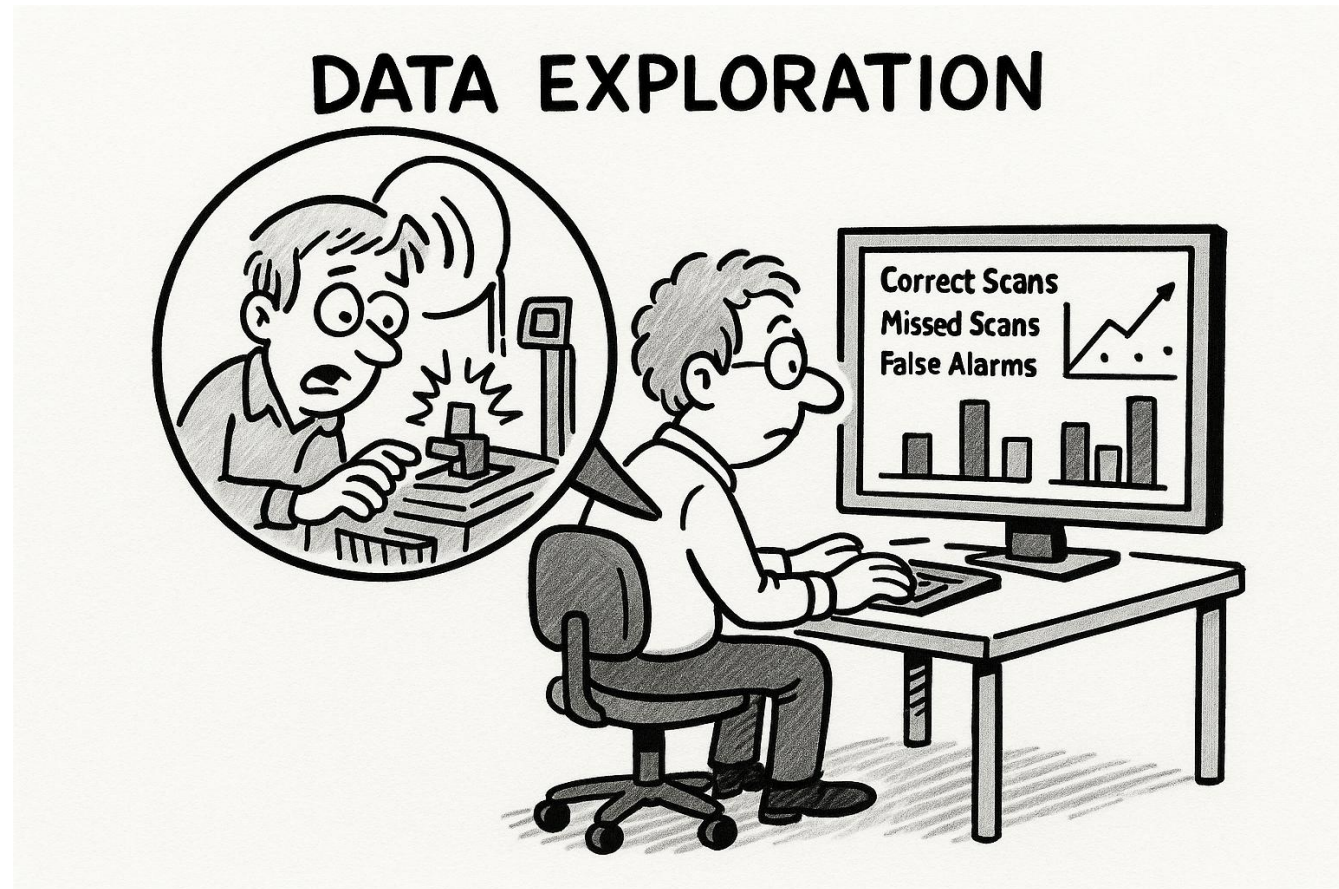


Zweiter Meilenstein - Datenbereitstellung

Verlustprävention an
Selbstbedienungskassen
im Einzelhandel

*Durchgeführt durch die
Retail Data Mining GmbH*

**Das Projekt-
Team bei der
Arbeit...**



1. Vorbemerkungen

Vorbemerkungen

- Vorbehalt: Methodische Anpassungen bei **neuen** Erkenntnissen
- Keine **negativen** Schadensfälle enthalten
- Annahme: Konformität mit **KassenSichV** (TSE) → Datenbasis ist vertrauenswürdig
- Begriffsdefinition: „line“ entspricht „Position“

2.Datenmanagment

Datenmanagement

- **Dateiformate:** .parquet für große Transaktionen/Lines, .csv für Stammdaten
- **Speicherung und Versionierung:** lokale Ablage, passwortgeschützte Einbindung in GitHub
- **Datenschutz:** Es sind keine personenbezogenen Daten enthalten – DSGVO-konform
- **Skalierbarkeit:** Alle Schritte in Jupyter Notebooks dokumentiert und modular aufgebaut für spätere Automatisierung

3.Vorabanalyse

Vorabanalyse

- **Repräsentativität der gelabelten Daten Speicherung und Versionierung**
- **Analyse des Storno-Prozesses**
- **Kunden-Feedback**
- **Kamerasystem**
- **Plausibilitätsprüfungen**

Vorabanalyse - Repräsentativität

Fazit: gelabelte Daten sind repräsentativ für den gesamten Datensatz

hier: für numerische Merkmale auf Basis eines t-Tests

Numerische Spalten (t-Tests):

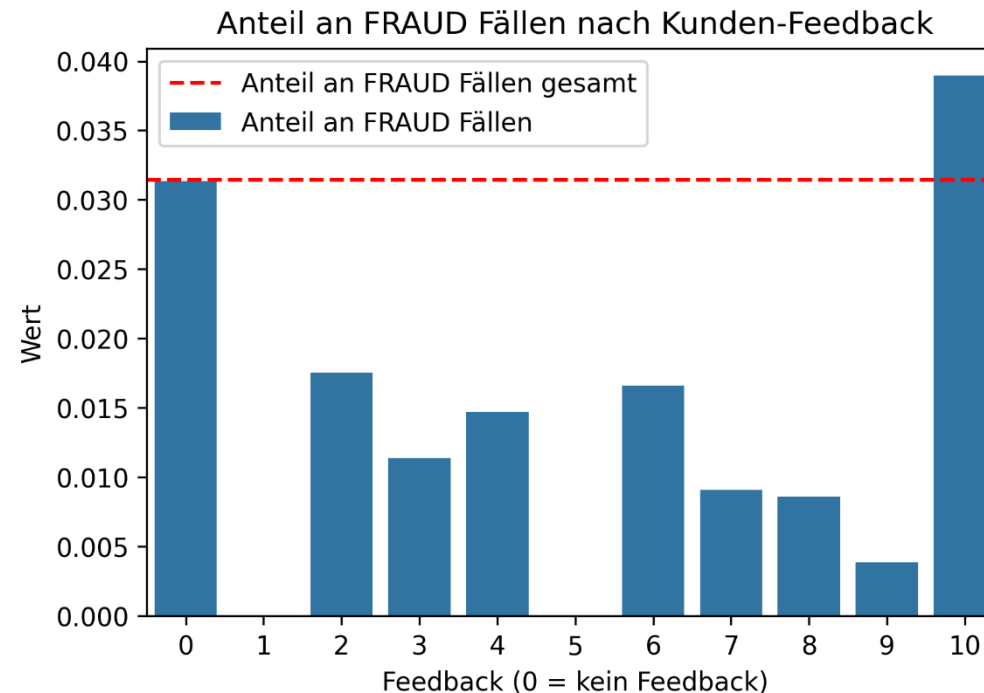
	Spalte	p-Wert	Mittelwert (labeled)	Mittelwert (unlabeled)	Std-Abw (labeled)	Std-Abw (unlabeled)
3	transaction_duration	0.185389	77.807475	77.541994	73.202614	72.895636
1	n_lines	0.355874	10.603607	10.575406	11.155176	11.101239
2	customer_feedback	0.671868	9.326005	9.318636	1.699571	1.715356
0	total_amount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

Vorabanalyse – Stornoprozess

- 500 stornierte Transaktionen mit `sales_price = 0` → alle als FRAUD klassifiziert
 - von Kamera als nicht-gescannte Artikel erkannt
 - weitere Behandlung dieser Fälle noch zu klären
- Weitere Auffälligkeiten im Zusammenhang mit Stornierungen → noch in Klärung

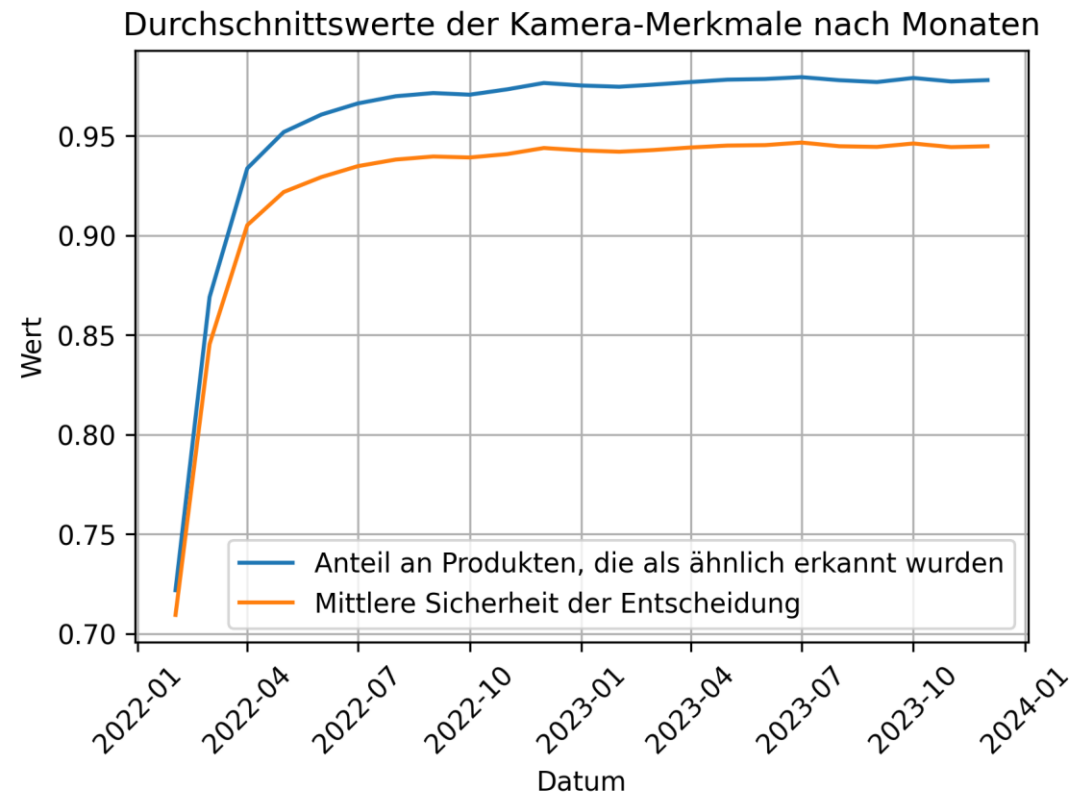
Vorabanalyse – Kundenfeedback

Bei FRAUD sind niedrige Kundenbewertungen unterrepräsentiert



Vorabanalyse – Kamerasystem

Lernkurve des Kamerasystems



Vorabanalyse-Plausibilitätsprüfungen

- Konsistenz von sales_price → NEIN
- Konsistenz von Summe sales_price und total_amount → NEIN
- Camera_certainty $\in [0,1]$ → JA
- Timestamp in Transaktionszeitraum → JA
- Damage > 0 nur bei label=FRAUD → JA
- Sales_price = 0,00 € → 100 % FRAUD

5. Daten- Transformation

Datentransformation

- Join über transaction_id, product_id, store_id → insgesamt 1,6 Mio. gelabelte Positionen
- Berechnete Merkmale: Dauer, Uhrzeit, Monat, Wochentag, Soll-Preis...
- Behandlung fehlender Werte

Datentransformation – Join-Strategie

Datentransformation – berechnete Merkmale

- Transaktionsdauer
- Uhrzeit, Wochentag, Monat
- Abstand zwischen zwei aufeinanderfolgenden Scanvorgängen
- Abstand Transaktionsstart zum ersten Scan / letzter Scan zum Transaktionsende
- Kalkulierter sales_price
- Kalkulierter total_amount

Datentransformation – fehlende Werte

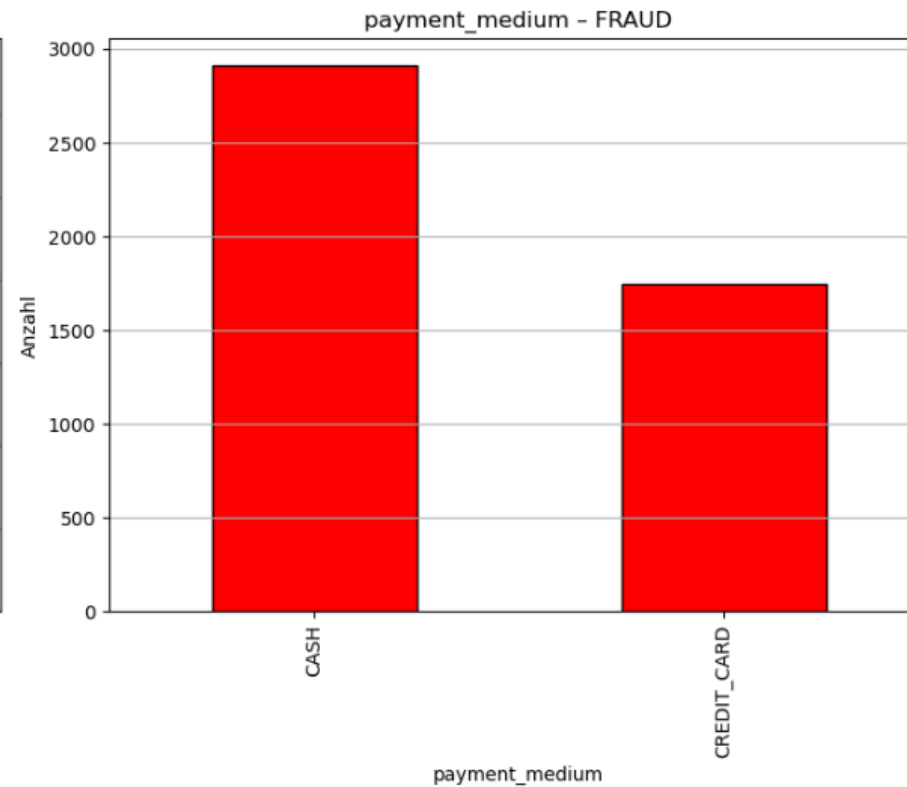
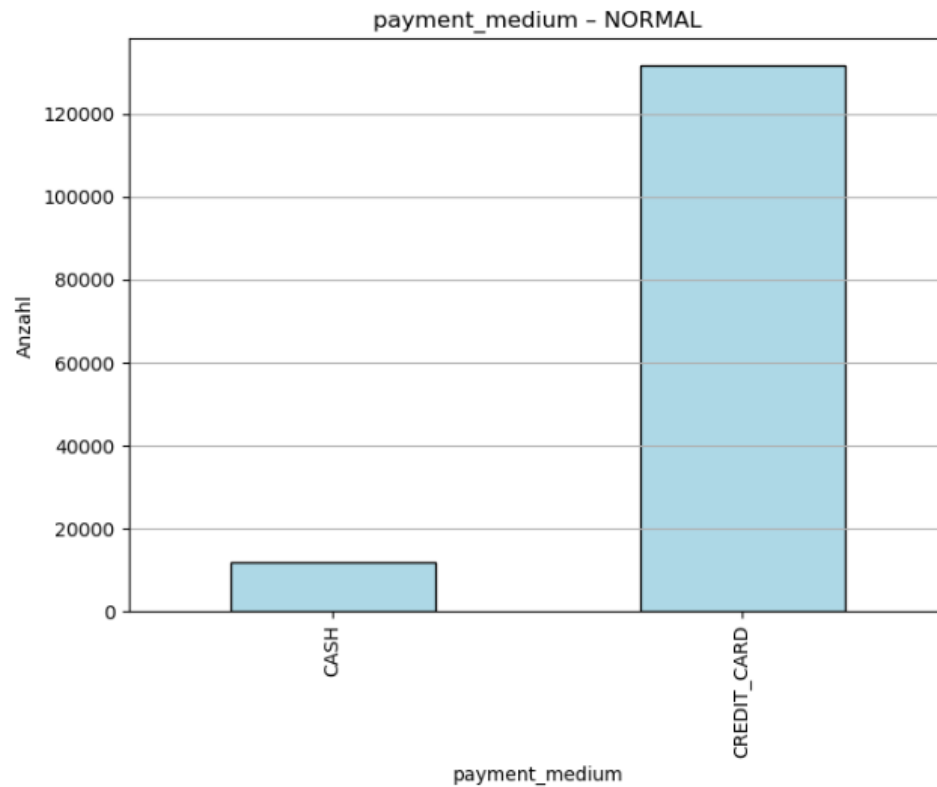
Datentransformation – fehlende Werte

5. Explorative Analyse

Explorative Analyse (EDA)

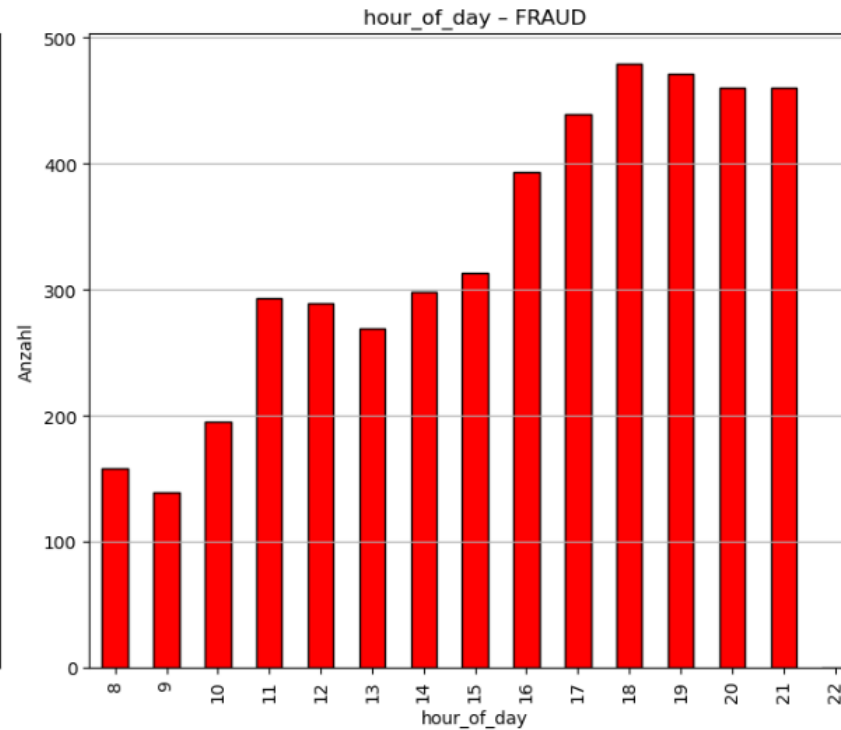
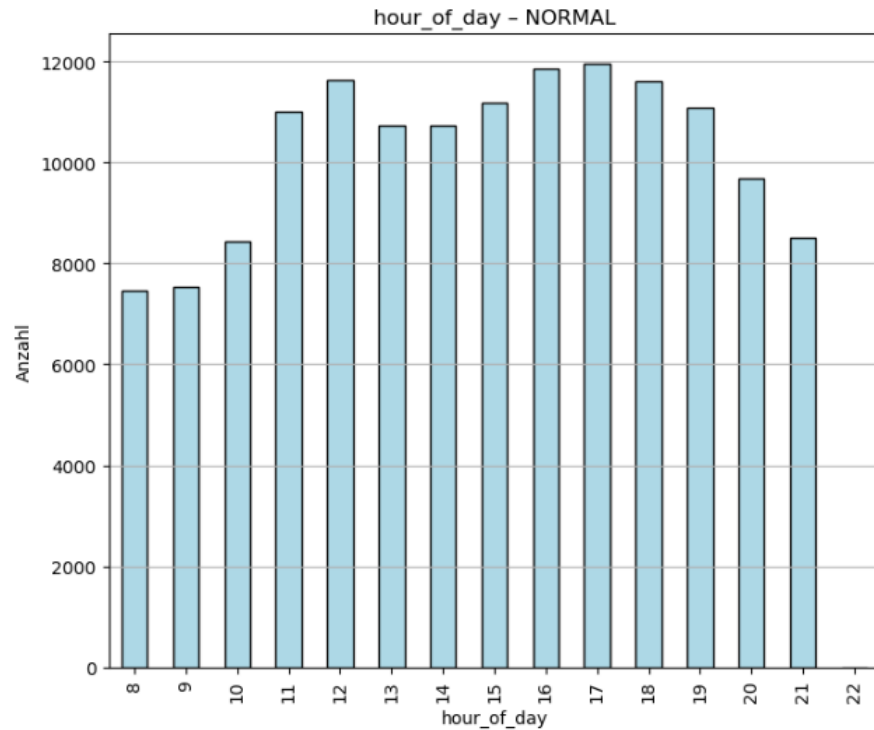
EDA- kategoriale Attribute

Payment_medium



EDA- kategoriale Attribute

Hour of day



Fazit

- Datengrundlage ist geeignet für Modellaufbau
- Relevante Merkmale für Klassifikation identifiziert
- Plausibilitäten geprüft
- Nächster Schritt: Modellierung
- Bewertungsfunktion → in Abstimmung mit Wertkauf GmbH



**Vielen Dank für Ihre
Aufmerksamkeit!**

Fragen & Anregungen?