2. Meilenstein - Datenbereitstellung

Verlustprävention an Selbstbedienungskassen im Einzelhandel

Durchgeführt durch die Retail Data Mining GmbH

RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik

Das Projekt-Team bei der Arbeit...

RDM
Retail Data Mining GmbH

5. Vorabanalyse

Vorabanalyse

- Vorbehalt: Methodische Anpassungen bei **neuen** Erkenntnissen
- Keine **negativen** Schadensfälle enthalten
- Annahme: Konformität mit **KassenSichV** (TSE) → Datenbasis ist vertrauenswürdig
- Verwendung **gelabelter** Daten: statistisch repräsentativ

RDM
Retail Data Mining GmbH

Vorabanalyse- labeled vs. unlabeled

gelabelte Daten sind repräsentativ

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

RDM
Retail Data Mining GmbH

5.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

RDM
Retail Data Mining GmbH

5. Transformation

Datentransformation

- Join über transaction_id, product_id, store_id
- Berechnete Merkmale: Dauer, Uhrzeit, Wochentag, Verkaufspreis
- Plausibilitätsprüfungen: rechnerische Richtigkeit, Zeitstempel innerhalb Transaktion

RDM
Retail Data Mining GmbH

Datentransformation

Datentransformation

Datentransformation

5. Explorative Analyse

Explorative Analyse (EDA)

- Nur **gelabelte** Transaktionen analysiert
- **Numerische** und **kategoriale** Attribute
- Vergleich FRAUD vs NORMAL über Histogramme / Boxplots
- Signifikante und relevante Unterschiede in :
  payment_medium, sales_price_difference und sales_price = 0

RDM
Retail Data Mining GmbH

EDA-Plausibilitätsprüfungen

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

RDM
Retail Data Mining GmbH

EDA-Plausibilitätsprüfungen

Vergleich rechnerischer und tatsächlicher Betrag
→ unberechtigter Rabatt

transaction_id	total_amount	sales_sum_per_transaction	damage	difference
001ee2b1-a10c-4577-9f29-74a510df2f98	31.81	37.93	6.12	-6.12
00207310-b2be-4691-ba8f-cd54c3db89c5	197.26	199.65	2.39	-2.39
003e9f5e-795c-4251-956f-b173f3bf01f3	9.94	13.70	3.76	-3.76
0045f1b7-f6c5-4e21-ba09-7334939e464e	93.88	104.39	10.51	-10.51
00a36958-159a-4f66-9861-63df5a27cbb3	55.34	60.85	5.51	-5.51

RDM
Retail Data Mining GmbH

EDA-Plausibilitätsprüfungen

Transaktions-Lines ohne Verkaufspreis → 100 % Fraud

```
sp_null = merged[merged.sales_price == 0.00]
```

```
sp_null.label.value_counts()
```

```
label
FRAUD    500
Name: count, dtype: int64
```

```
sp_null.was_voided.value_counts()
```

```
was_voided
True    500
Name: count, dtype: int64
```

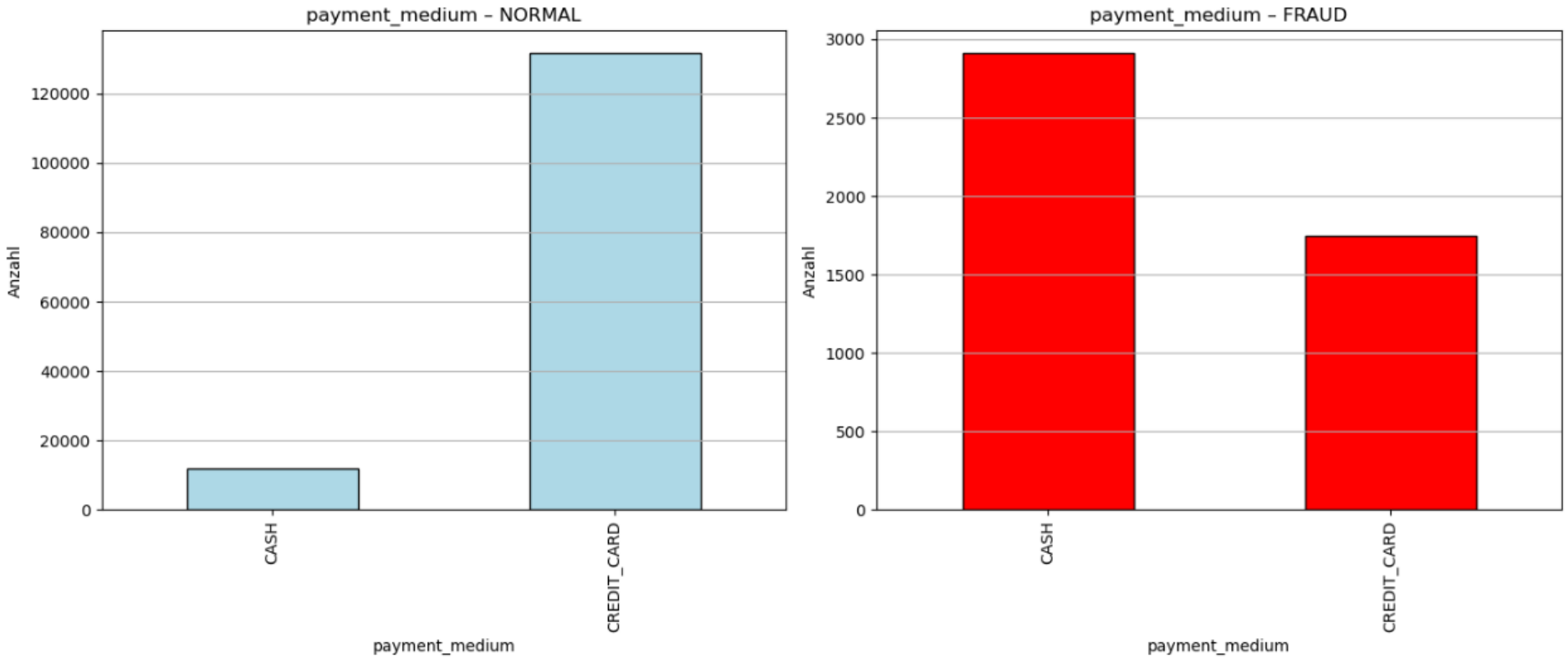
```
sp_null.damage.sum()
```

```
7474.9400000000005
```

RDM
Retail Data Mining GmbH

EDA- kategoriale Attribute

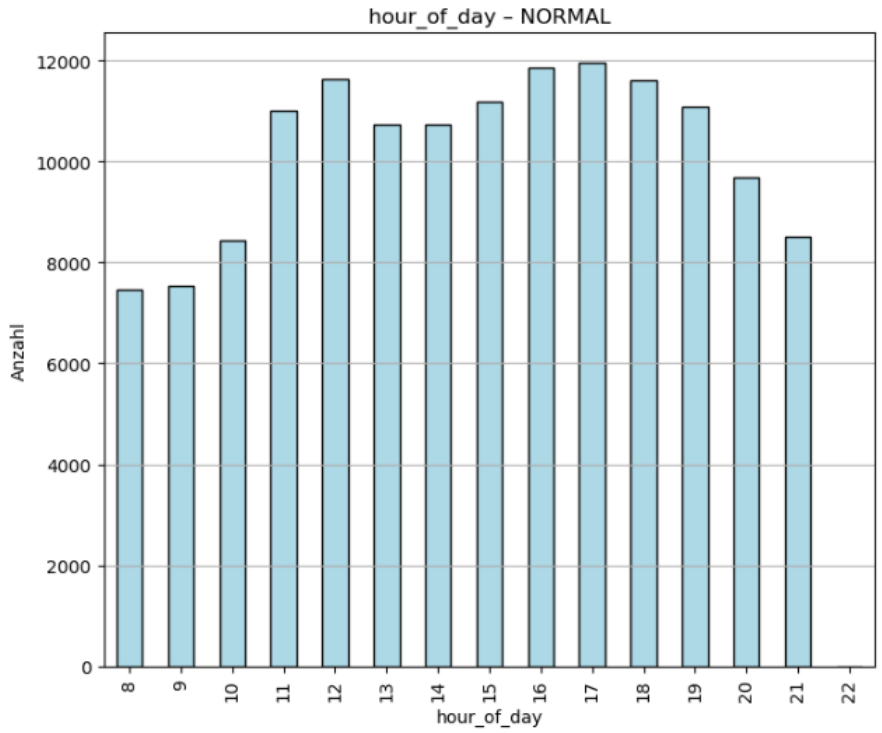
Payment_medium

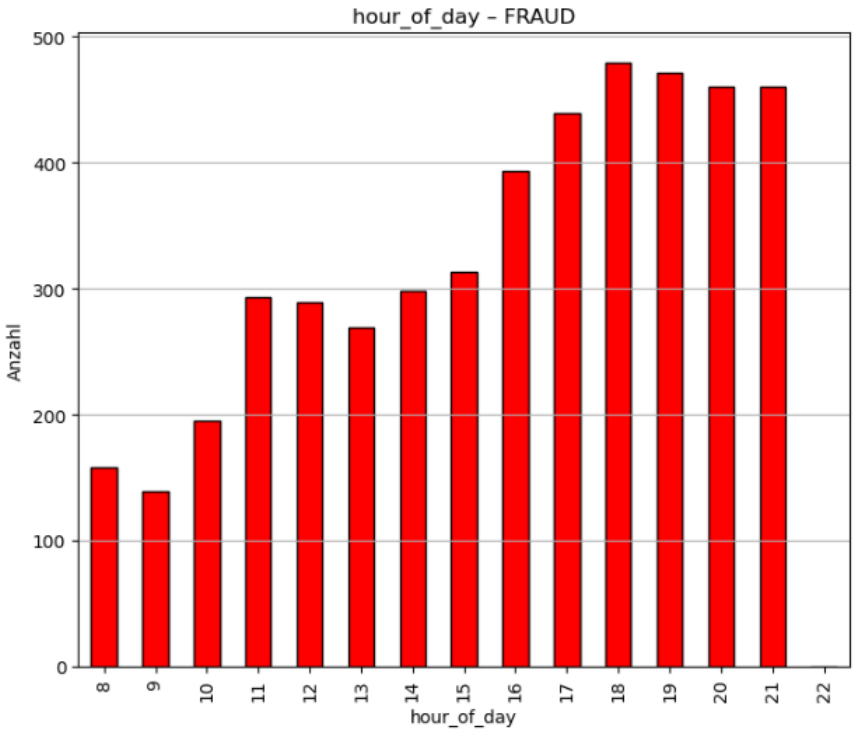


RDM
Retail Data Mining GmbH

EDA- kategoriale Attribute

Hour of day





RDM
Retail Data Mining GmbH

EDA- kategoriale Attribute

Absolute Häufigkeiten in 'weekday':

label	FRAUD	NORMAL
weekday		
Friday	1080	32857
Monday	768	23949
Saturday	1377	40766
Thursday	591	17136
Tuesday	425	14744
Wednesday	415	13917

Chi-Quadrat-Test für 'weekday':
  Chi² = 13.21, p-Wert = 0.0215, Freiheitsgrade = 5
  → Ergebnis ist signifikant (p < 0.05)

Absolute Häufigkeiten in 'hour_of_day':

label	FRAUD	NORMAL
hour_of_day		
8	158	7454
9	139	7535
10	195	8431
11	293	11004
12	289	11624
13	269	10740
14	298	10728
15	313	11188
16	393	11843
17	439	11946
18	479	11594
19	471	11091
20	460	9676
21	460	8514
22	0	1

Chi-Quadrat-Test für 'hour_of_day':
  Chi² = 405.08, p-Wert = 0.0000, Freiheitsgrade = 14
  → Ergebnis ist signifikant (p < 0.05)

RDM
Retail Data Mining GmbH

EDA - kategoriale Attribute

Produktgruppe

=== Analyse der Schadensverteilung in Kategorie: category ===

	Anzahl aller Transaktionen	Anteil aller Transaktionen (%)	Schadenssumme	Anteil Schaden (%)
FRUITS_VEGETABLES_PIECES	743041	46.98	201538.29	46.16
BEVERAGES	412908	26.11	109432.35	25.07
DAIRY	135168	8.55	43636.10	9.99
FROZEN_GOODS	129638	8.20	32131.06	7.36
CONVENIENCE	62157	3.93	22377.22	5.13
LONG_SHELF_LIFE	65354	4.13	14598.38	3.34
SNACKS	3470	0.22	4832.04	1.11
PERSONAL_CARE	14487	0.92	3806.01	0.87
HOUSEHOLD	6113	0.39	1659.94	0.38
FRUITS_VEGETABLES	3791	0.24	1206.86	0.28
BAKERY	2023	0.13	542.55	0.12
ALCOHOL	2556	0.16	441.06	0.10
TOBACCO	844	0.05	385.43	0.09

RDM
Retail Data Mining GmbH

EDA- numerische Attribute

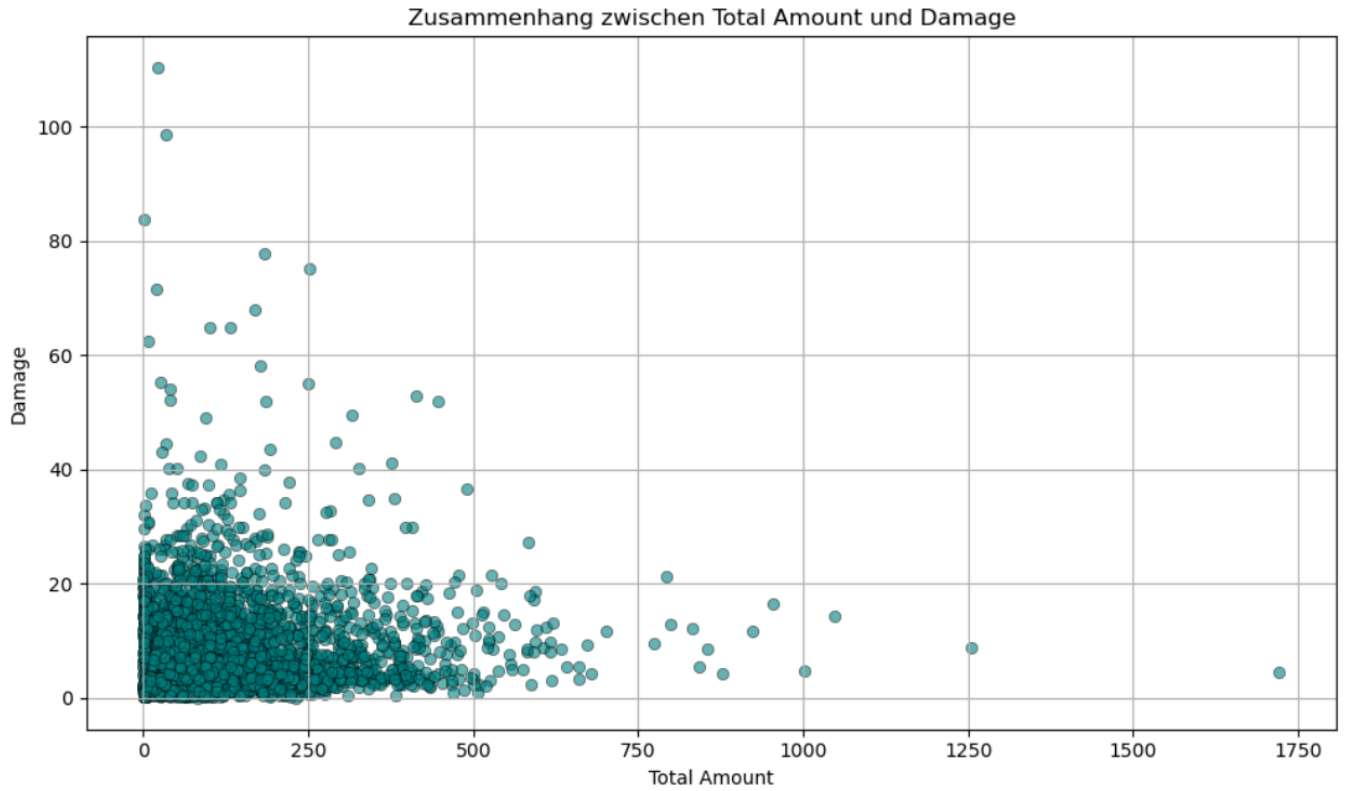
```
t-Test für 'duration_per_line':
  t-Wert = -4.78, p-Wert = 0.0000
  → Ergebnis ist signifikant (p < 0.05)
------------------------------------------------------------
```

Zusammenfassung der t-Tests für numerische Merkmale:

	Spalte	t-Wert	p-Wert	Signifikant (p < 0.05)
0	total_amount	-0.85	0.3948	Nein
1	n_lines	-4.41	0.0000	Ja
2	customer_feedback	-10.03	0.0000	Ja
3	transaction_duration	-5.23	0.0000	Ja
4	duration_per_line	-4.78	0.0000	Ja

RDM
Retail Data Mining GmbH

EDA- numerische Attribute



RDM
Retail Data Mining GmbH

Fazit

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

RDM
Retail Data Mining GmbH

Vielen Dank für Ihre Aufmerksamkeit!

Fragen & Anregungen

RDM
Retail Data Mining GmbH

Matthias Bald | David Zurschmitten | Raphael Schaffarczik