DATENBEREITSTELLUNG

Verlustprävention an Selbstbedienungskassen im Einzelhandel

Projektgruppe: Raphael Schaffarczik, David Zurschmitten, Matthias Bald
Auftraggeber: Wertkauf GmbH

Abgabedatum: 20.05.2025

Dokumentation – Meilenstein 2: Datenbereitstellung

Vorbemerkungen

Der zweite Meilenstein im Rahmen des DASC-PM umfasst die **Datenaufbereitung**, das **Datenmanagement** sowie die **explorative Analyse** zur Vorbereitung der nachfolgenden Modellierungsschritte.

Methodische Anpassung bei neuen Erkenntnissen

Im Sinne eines iterativen und explorativen Data-Science-Prozesses behalten wir uns vor, Analysen, Annahmen und Modellierungsstrategien im weiteren Projektverlauf anzupassen, sofern sich durch neue Erkenntnisse oder zusätzliche Daten relevante Änderungen ergeben. Diese Flexibilität entspricht den Anforderungen an eine verantwortungsbewusste, datengestützte Entscheidungsfindung und steht im Einklang mit dem Vorgehensmodell DASC-PM.

Konformität mit der Kassensicherungsverordnung

Wir gehen im Rahmen dieses Projekts davon aus, dass die Kassensysteme der Wertkauf GmbH den Anforderungen der **Kassensicherungsverordnung (KassenSichV)** entsprechen. Insbesondere wird unterstellt, dass sämtliche Kassenaufzeichnungen durch eine **zertifizierte technische Sicherheitseinrichtung (TSE)** manipulationssicher protokolliert wurden und dass die uns zur Verfügung gestellten Daten konsistent sind.
Nach Auskunft der Gesellschaft können allerdings neben den bereits zur Verfügung gestellten Metadaten zu den Spalten der Dateien keine ergänzenden Dokumentationen, Auditberichte oder technische Nachweise bereitgestellt werden, auf die wir unsere Analysen zusätzlich stützen könnten.

Mögliche Verzerrung durch fehlende negative Schadensfälle

Im bereitgestellten Datensatz sind ausschließlich Transaktionen mit einem **positiven finanziellen Schaden** enthalten.
In der Realität ist jedoch davon auszugehen, dass auch gegenteilige Fälle auftreten – beispielsweise, wenn Kunden mehr Artikel bezahlen, als sie tatsächlich mitnehmen, etwa durch vergessene Ware im Geschäft oder unvollständige Stornierungen.
Diese „**negativen Schäden**“ sind in den verfügbaren Daten nicht enthalten, was zu einer gewissen Verzerrung der Modellbewertung führen kann, insbesondere im Hinblick auf Nettoverlustanalysen.

Begriffsdefinition

Der im Englischen verwendete Begriff *line* – im Sinne einzelner Einträge innerhalb einer Transaktion – wird im Folgenden mit *Position* (die Positionen einer Transaktion) bezeichnet.

Datenmanagement

Struktur und Handhabung

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

Vorabanalyse

Verwendung gelabelter Daten auf Basis statistischer Repräsentativität

Für sämtliche weitere Analysen und Modellierungen im Rahmen dieser Projektphase wurden ausschließlich die **gelabelte Transaktionen** berücksichtigt (mit label = FRAUD oder label = NORMAL).
Die Fokussierung auf diese Teilmenge ist gerechtfertigt, da mithilfe statistischer Tests (**Chi-Quadrat-Test** für kategoriale Merkmale und **t-Test** für numerische Merkmale) gezeigt wurde, dass die gelabelten Daten in ihrer Merkmalsverteilung **repräsentativ für den Gesamtdatensatz** sind.

Im Gegensatz dazu konnte für die **Testdaten** keine vollständige Repräsentativität gegenüber den Trainingsdaten festgestellt werden.

Diese Abweichung könnte durch **zeitliche Unterschiede** zwischen den Datensätzen oder z.B. durch den **Lernprozess des Kamerasystems** bedingt sein.

Differenz zwischen sales_price und errechnetem Nominalpreis

Im Rahmen einer rechnerischen Überprüfung wurde untersucht, ob der *sales_price* der einzelnen Positionen der Transaktionen mit dem nominal zu erwartenden Preis übereinstimmt. Letzterer ergibt sich aus dem Produkt von *price* und *pieces_or_weight*. Dabei wurden signifikante Abweichungen festgestellt, die in FRAUD-Fällen überproportional häufig auftreten. Diese Erkenntnis floss direkt in die Merkmalsauswahl ein (siehe Abschnitt Datentransformation) und wurde durch die Konstruktion entsprechender Merkmale berücksichtigt.

In Rücksprache mit der Wertkauf GmbH ergab sich eine plausible Erklärung für diesen Befund: Die betroffenen Fälle könnten auf die unrechtmäßige Anwendung von Standardrabatten –

insbesondere der 30%-Nachlass auf Ware kurz vor Ablauf des Mindesthaltbarkeitsdatums – zurückzuführen sein.

Seitens der Wertkauf GmbH wurde der Wunsch geäußert, den potenziellen finanziellen Schaden durch diese spezifische Betrugsmasche zu quantifizieren und zu untersuchen, inwieweit sich dieser durch technische Maßnahmen verringern lässt.

Dieser Fragestellung werden wir im weiteren Projektverlauf selbstverständlich nachgehen und die Ergebnisse der Wertkauf GmbH zur Verfügung stellen. Da diese Information jedoch erst nach Abschluss der Merkmalsauswahl und der Datentransformation verfügbar wurde, ist sie in der nachfolgenden Analyse noch nicht berücksichtigt.

Analyse des Storno-Prozesses

Im Rahmen der Analyse wurden 500 Transaktionen identifiziert, in denen mindestens eine stornierte Position enthalten ist, der *sales_price* exakt 0,00 € beträgt und das Merkmal *camera_product_similar* den Wert *false* aufweist. Alle diese Transaktionen sind als *FRAUD* klassifiziert.

Nach Rücksprache mit der Wertkauf GmbH hat sich erwiesen, dass das Kamerasystem in der Lage ist, eigenständig Positionen in eine Transaktion einzufügen, wenn es einen Artikel erkennt, der nicht gescannt wurde. Damit diese erkannten, aber nicht registrierten Produkte nicht auf dem Kassenbon erscheinen, werden sie vom System als *storniert* gekennzeichnet. Im Unterschied zu regulär stornierten Positionen weisen diese jedoch einen *sales_price* von 0,00 € auf.

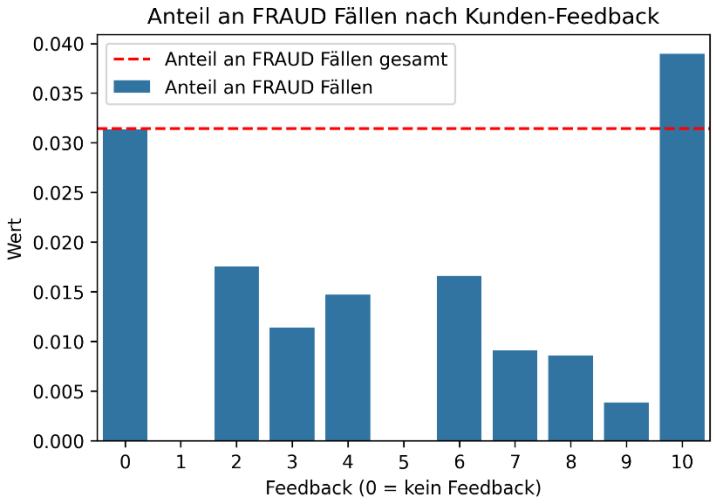
Diese Betrugsfälle lassen sich durch eine gezielte Abfrage eindeutig identifizieren. Allerdings stellt sich die Frage, ob in solchen Situationen nicht bereits eine Intervention durch das Personal erfolgt und der Sachverhalt damit als erledigt betrachtet werden kann. Eine entsprechende Rückmeldung seitens der Wertkauf GmbH steht hierzu noch aus.

Da diese Information erst zu einem späten Zeitpunkt im Projektverlauf zur Verfügung stand, konnte sie bislang nicht in die Merkmalsauswahl und Analyse integriert werden. Eine nachträgliche Einbindung ist jedoch vorgesehen.

Weitere Auffälligkeiten und offene Fragen im Zusammenhang mit der Spalte *was_voided* wurden ebenfalls mit der Wertkauf GmbH abgestimmt. Die daraus gewonnenen Erkenntnisse erfordern eine weiterführende Analyse und Bewertung. Sofern sich daraus relevante Muster oder Zusammenhänge ergeben, werden diese in zukünftige Modellversionen einfließen.

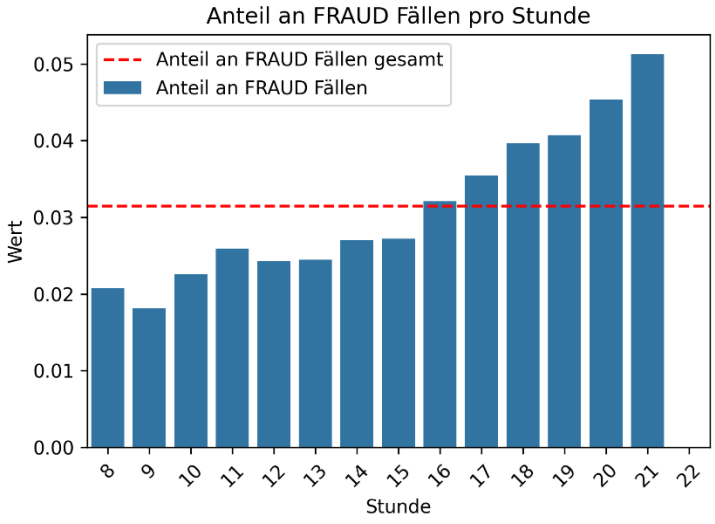
Kunden-Feedback

Die Spalte *customer_feedback* weist mit Abstand die höchste Anzahl fehlender Werte im Datensatz auf – lediglich in 7,6 % der Fälle liegt ein Feedback-Wert vor. Zudem ist der Mittelwert mit 9,3 außergewöhnlich hoch. Wie der folgende Plot verdeutlicht, tritt insbesondere der Wert 10 in *FRAUD*-Fällen überproportional häufig auf, während alle anderen Feedback-Werte deutlich unterrepräsentiert sind.



Relevanz der Tageszeit

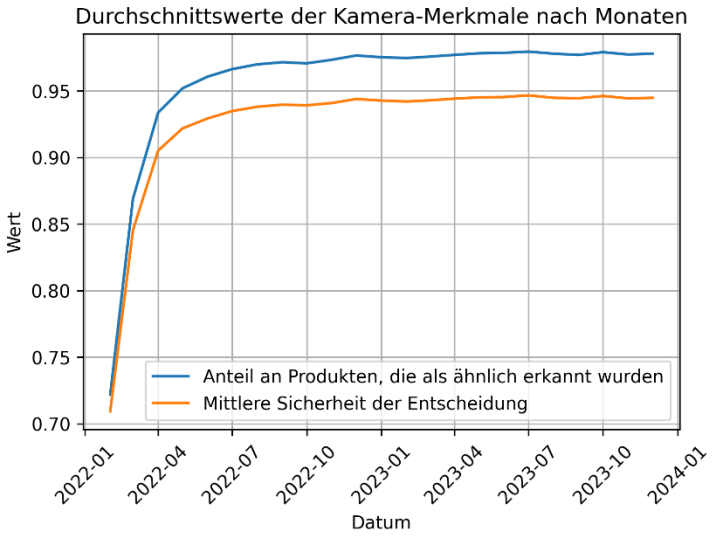
Eine Analyse der Transaktionszeitpunkte in Bezug auf periodische Kategorien wie Wochentag, Monat und Stunde zeigte, dass insbesondere die Tagesstunde einen deutlichen Einfluss auf das Auftreten von *FRAUD*-Fällen hat. Die Anteile betrügerischer Transaktionen nimmt im Tagesverlauf kontinuierlich zu.



Kamerasystem

Wie auf dem folgenden Plot gut zu sehen ist, war das Kamerasystem in der Anfangsphase nach Einführung der Selbstbedienungskassen noch wenig zuverlässig und befand sich offensichtlich in einer Trainings- bzw. Kalibrierungsphase. Erst ab Juli 2022 stabilisieren sich sowohl die Anzahl der von der Kamera erkannten Artikel als auch die durchschnittliche Vorhersagesicherheit langsam. Die durchschnittlichen Werte in *camera_certainty* steigen kontinuierlich an und erreichen einen Wert von knapp 0,95. In rund 98 % der Fälle erkennt das System ab diesem Zeitpunkt das gescannte Produkt als ähnlich – ein Indikator für die zunehmende Reife des Modells.

Auch bei neu eingeführten Produkten sind Vorhersagen des Kamerasystems erwartungsgemäß unzuverlässig. Diesem Umstand sollte bei der Auswahl der Merkmale für die weitere Analyse Rechnung getragen werden.



Datentransformation

Ausgehend von den Ergebnissen der Vorabanalyse wurde ein konsolidierter Datensatz auf Basis der vier ursprünglichen Datenquellen erstellt. Dabei erfolgte eine gezielte Auswahl relevanter Merkmale sowie die Generierung abgeleiteter Variablen, um die Aussagekraft der Daten für die anschließende statistische Analyse zu erhöhen.

Die Zusammenführung der Daten erfolgte nach einem strukturierten Vorgehen: Transaktionsdaten wurden um standortbezogene Informationen aus der Datei *stores.csv* angereichert, während die einzelnen Positionen durch produktspezifische Merkmale aus *products.csv* ergänzt wurden. Aus den daraus resultierenden Zwischenschritten wurde ein finaler Datensatz erstellt, der jede Transaktion in einer Zeile abbildet. Dabei wurden die für die Analyse relevanten Informationen aus den Positionen auf Transaktionsebene aggregiert.

Bei der Auswahl der Merkmale orientierten wir uns sowohl an den Ergebnissen der Vorabanalyse als auch an der im Team vorhandenen fachlichen Erfahrung hinsichtlich der Trennschärfe potenzieller Variablen. Ziel war es, Merkmale zu identifizieren, die mit hoher Wahrscheinlichkeit relevante Unterschiede im Datensatz abbilden können. So wurde beispielsweise angenommen, dass der maximale Artikelpreis innerhalb einer Transaktion eine höhere Aussagekraft besitzt als der minimale.

Grundsätzlich verfolgten wir bei der Merkmalsauswahl einen eher inklusiven Ansatz: Auch Merkmale, deren Relevanz zunächst unklar erschien, wurden vorerst in den Datensatz aufgenommen, sofern sie potenziell analytischen Mehrwert bieten konnten. Die anschließende explorative Datenanalyse sollte dann empirisch zeigen, welche dieser Variablen tatsächlich zur Erklärung oder Trennung von Datenmustern beitragen.

Im Folgenden werden die ausgewählten Merkmale kurz erläutert.

1. Merkmalselektion

Aus dem kombinierten Datensatz aus Transaktionen und Filialen wurden folgende Spalten unverändert übernommen:

- *cash_desk*
- *total_amount*
- *n_lines*
- *payment_medium*
- *location*
- *urbanization*

Zusätzlich wurden folgende Merkmale erzeugt:

- *has_feedback, feedback_low*, *feedback_middle*, *feedback_high, feedback_top*
- *transaction_duration*
- *month, weekday*
- *daytime, hour, hour_categorical*
- days_since_sco_introduction

Kundenfeedback:

Wie bereits in der Vorabanalyse aufgezeigt, weist die Spalte *customer_feedback* eine geringe Befüllung auf, wobei der Wert 10 stark überrepräsentiert ist. Um diesen Gegebenheiten angemessen zu begegnen, wurden aus dieser Spalte insgesamt fünf neue Merkmale abgeleitet: Zum einen ein binäres Merkmal, das angibt, ob überhaupt ein Feedback vorliegt (*has_feedback*), zum anderen vier zusätzliche binäre Variablen zur Repräsentation verschiedener Bewertungsniveaus: *feedback_low*, *feedback_middle*, *feedback_high* und *feedback_top*.

Diese Kategorisierung ermöglicht es, die potenzielle Aussagekraft der wenigen vorhandenen Bewertungen zu nutzen, ohne fehlende Werte durch künstliche Imputation ersetzen zu müssen. Zugleich kann auch der Umstand, **dass** ein Kunde Feedback abgegeben hat, als mögliches trennscharfes Merkmal in die Analyse einfließen.

Da der Wert 10 besonders häufig auftritt, wurde diesem mit *feedback_top* ein eigenes Merkmal zugewiesen. Die übrigen Bewertungen wurden in drei Gruppen zusammengefasst: *feedback_low* (Werte 1–3), *feedback_middle* (4–6) und *feedback_high* (7–9). Dieses Vorgehen reduziert die Dimensionalität und wahrt dennoch die potenzielle Aussagekraft der Rückmeldungen.

Zeitpunkt der Transaktion:

Aus den Zeitstempeln *transaction_start* und *transaction_end* wurden mehrere Merkmale abgeleitet. Zum einen wurde die Transaktionsdauer in Sekunden berechnet (*transaction_duration_seconds*), zum anderen wurden periodische Zeitinformationen extrahiert: der Wochentag, der Monat (*weekday, month*) sowie die Tageszeit in drei unterschiedlichen Darstellungsformen (*daytime, hour, hour_categorical*). Zum einen eine Einteilung in vier grobe

Tageszeitkategorien, eine kategoriale Einteilung nach vollen Stunden sowie eine numerische Darstellung der Stunde.

Die verschiedenen Varianten der Tageszeitdarstellung wurden bewusst parallel berücksichtigt, um im weiteren Analyseverlauf vergleichen zu können, welche Repräsentation den größten Beitrag zur Modellgüte liefert. Bereits in der Vorabanalyse deuteten sich deutliche Unterschiede im Auftreten von Betrugsfällen (*FRAUD*) in Abhängigkeit von der Tageszeit an, weshalb diesem Merkmal besondere Aufmerksamkeit geschenkt wurde.

Zur Berücksichtigung potenzieller Lerneffekte sowie möglicher Veränderungen im Nutzungs- oder Betrugsverhalten im zeitlichen Verlauf wurde das Merkmal *days_since_sco_introduction* eingeführt, das die Anzahl der Tage seit der Einführung des Self-Checkout-Systems in der jeweiligen Filiale abbildet.

Merkmale aus den um die Produktinformation angereicherten Lines:

Da eine Transaktion jeweils mehrere Positionen (*Lines*) umfassen kann, wurden die in diesen enthaltenen Informationen auf Transaktionsebene aggregiert. Ziel war es, aus den einzelnen Produkteigenschaften summarische Merkmale zu erzeugen, die das Einkaufsverhalten pro Transaktion abbilden.

Numerische Aggregationen:

- *popularity_max*, *popularity_min*: Minimaler und maximaler Wert der Produktsäule *popularity* innerhalb einer Transaktion.
- *max_product_price*: Höchster Produktpreis unter allen Positionen einer Transaktion.

Binärmerkmale und Zählwerte für besondere Produkttypen:
Für die folgenden Spalten wurden jeweils zwei Merkmale erzeugt:

1. Ein binäres Merkmal, das angibt, ob mindestens eine Position mit dem entsprechenden Kennzeichen vorhanden ist (*has_...*),
2. sowie ein numerisches Merkmal mit der Anzahl dieser Positionen (*n_...*):

- *has_voided*, *n_voided*: Enthält die Transaktion stornierte Positionen (*was_voided*).
- *has_age_restricted*, *n_age_restricted*: Enthält die Transaktion altersbeschränkte Produkte (*age_restricted*).
- *has_sold_by_weight*, *n_sold_by_weight*: Enthält die Transaktion Produkte, die nach Gewicht verkauft wurden (*sold_by_weight*).

Kamerabasierte Abweichungserkennung:
Besondere Aufmerksamkeit wurde den Fällen gewidmet, in denen das Kamerasystem eine Abweichung zwischen gescanntem und erfasstem Produkt identifiziert hat. Hier wurden zwei Merkmale definiert:

- *has_camera_detected_wrong_product*: Gibt an, ob in der Transaktion mindestens eine Position mit erkannter Abweichung enthalten ist.
- *has_camera_detected_wrong_product_high_certainty*: Zusätzliches Merkmal, das nur Fälle mit einer *camera_certainty* über einem definierten Schwellenwert (0,8) berücksichtigt, da das System in den ersten Monaten noch trainiert wurde und teils unzuverlässig war.

Produktkategorien:
Für jede vorhandene Produktkategorie wurde ein binäres Merkmal erstellt, das angibt, ob mindestens eine Position dieser Kategorie in der Transaktion enthalten ist (*has_category_X*). Um mit fehlenden *product_id*-Werten umzugehen, wurde zusätzlich die Ersatzkategorie *has_missing* eingeführt, die das Vorhandensein solcher Positionen kennzeichnet.

Weitere abgeleitete Merkmale

In der Vorabanalyse zeigte sich ein auffälliges Muster: Transaktionen, bei denen der errechnete Nominalpreis nicht mit dem tatsächlich gezahlten Preis übereinstimmt, wiesen signifikant häufiger einen *FRAUD*-Fall auf. Auf Basis dieser Beobachtung wurden zwei Merkmale entwickelt:

- **calculated_price_difference**: Differenz zwischen dem errechneten Nominalpreis (basierend auf *price* × *pieces_or_weight*) und dem tatsächlichen Verkaufspreis (*sales_price*), summiert über alle Positionen einer Transaktion.
- **has_positive_price_difference**: Binäres Merkmal, das angibt, ob die berechnete Differenz positiv ist (d. h. > 0,01). Dieses Merkmal soll erfassen, ob Produkte systematisch günstiger verkauft wurden, als es dem Nominalpreis entspricht.

Temporale Merkmale aus Positionsdaten:

Aus den Zeitstempeln der einzelnen Positionen sowie den Transaktionszeitpunkten (*transaction_start*, *transaction_end*) wurden folgende Merkmale berechnet:

- **mean_time_between_scans**: Durchschnittlicher zeitlicher Abstand zwischen aufeinanderfolgenden Scans.
- **max_time_between_scans**: Längster zeitlicher Abstand zwischen zwei aufeinanderfolgenden Scans.
- **time_to_first_scan**: Zeitspanne vom Transaktionsstart bis zum ersten Scan.
- **time_from_last_scan_to_end**: Zeitspanne zwischen dem letzten Scan und dem offiziellen Transaktionsende.

Diese Merkmale dienen insbesondere der Erkennung ungewöhnlicher zeitlicher Muster, wie z. B. längerer Unterbrechungen oder ungewöhnlich schneller Scanfolgen, die potenziell auf betrügerisches Verhalten hindeuten könnten.

2. Behandlung fehlender Werte

Der auf diese Weise erzeugte Datensatz enthält an einzelnen Stellen fehlende Werte.

Bei Transaktionen, die nur eine einzige Position umfassen, lassen sich die Merkmale *mean_time_between_scans* und *max_time_between_scans* definitionsgemäß nicht berechnen, da hierfür mindestens zwei Zeitstempel benötigt werden. Um diese Lücken konsistent zu füllen und gleichzeitig Verzerrungen zu vermeiden, wurde entschieden, den Mittelwert dieser Merkmale innerhalb der jeweiligen Zielkategorie (*FRAUD*, *NORMAL*, *UNKNOWN*) zu verwenden. Diese kategorial differenzierte Imputation berücksichtigt potenziell systematische Unterschiede im Scanverhalten zwischen betrügerischen und regulären Transaktionen und minimiert damit das Risiko, strukturelle Muster zu verfälschen.

Auch in den Spalten *has_camera_detected_wrong_product* und *has_camera_detected_wrong_product_high_certainty* treten vereinzelt fehlende Werte auf,

vermutlich verursacht durch temporäre Ausfälle oder technische Einschränkungen des Kamerasystems. In diesen Fällen wurde der Modus – also der am häufigsten vorkommende Wert – innerhalb der jeweiligen Zielkategorie verwendet.

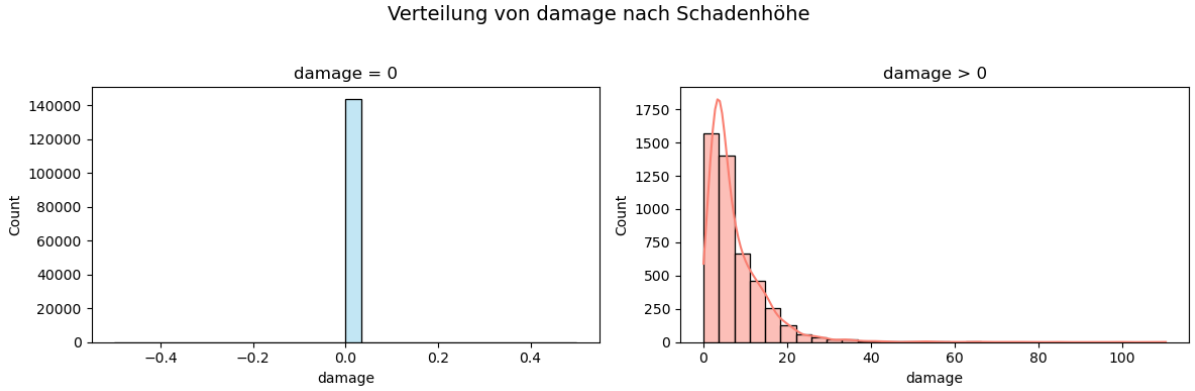
Einige wenige Zeilen im Datensatz weisen eine große Anzahl fehlender Werte auf. Diese Fälle sind darauf zurückzuführen, dass die entsprechenden Transaktionen keinerlei Positionen enthalten – womöglich aufgrund technischer Fehler bei der Datenerfassung oder abgebrochener Vorgänge. In den gelabelten Daten betrifft dies lediglich einen einzelnen Fall. Aufgrund der geringen Anzahl sind diese Zeilen für die statistische Analyse nicht relevant und wurde daher entfernt.

Explorative Datenanalyse

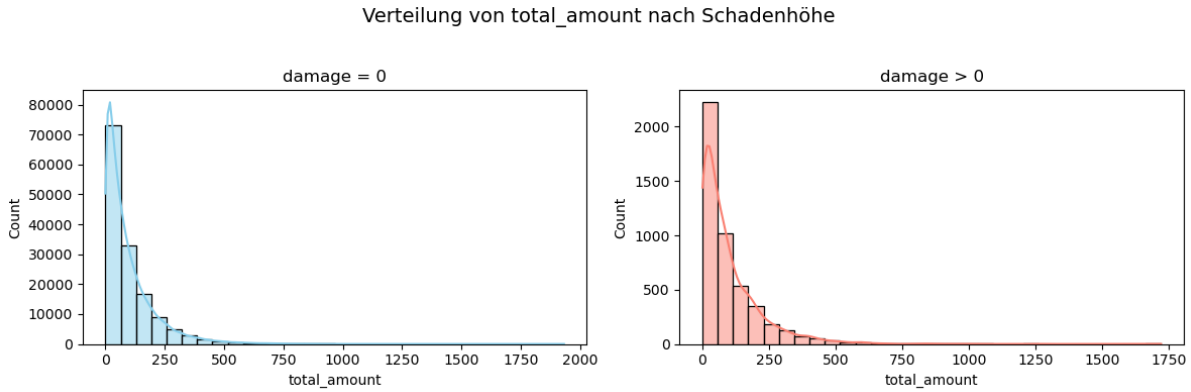
1. Verteilungsanalyse und Ausreißer numerischer Attribute

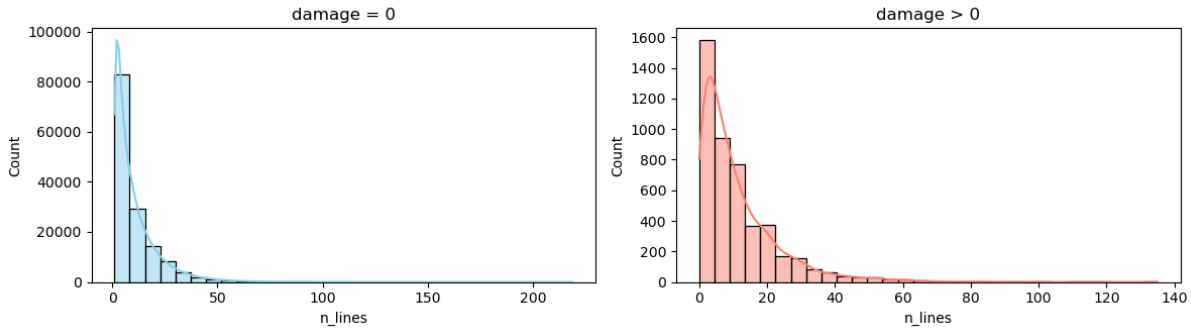
Wir betrachten nur noch die klassifizierten Daten („FRAUD“ bzw. „NORMAL“) mir ihren Pendant Schaden größer bzw. gleich Null. Gewisse Attribute zeigen deutliche Abweichung in ihrer Verteilung im Vergleich zwischen Schadensfall und nicht-Schadensfall.

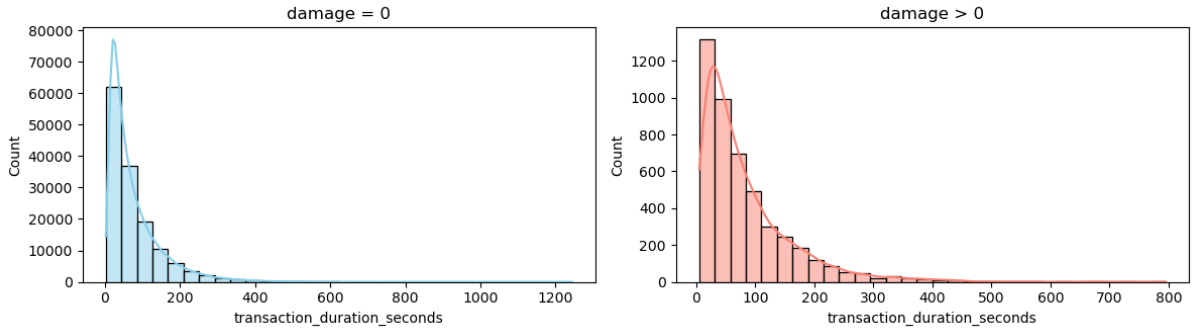
Die Zielvariable damage ist erwartungsgemäß bei der Mehrheit der Transaktionen exakt 0, da es sich hier um keine fehlerhafte oder betrügerische Transaktion handelt, entsprechend alle Artikel korrekt gescannt und verbucht wurden. Nur bei damage > 0 zeigt sich eine stark rechtsschiefe Verteilung mit wenigen, aber teils erheblichen Schadensbeträgen.



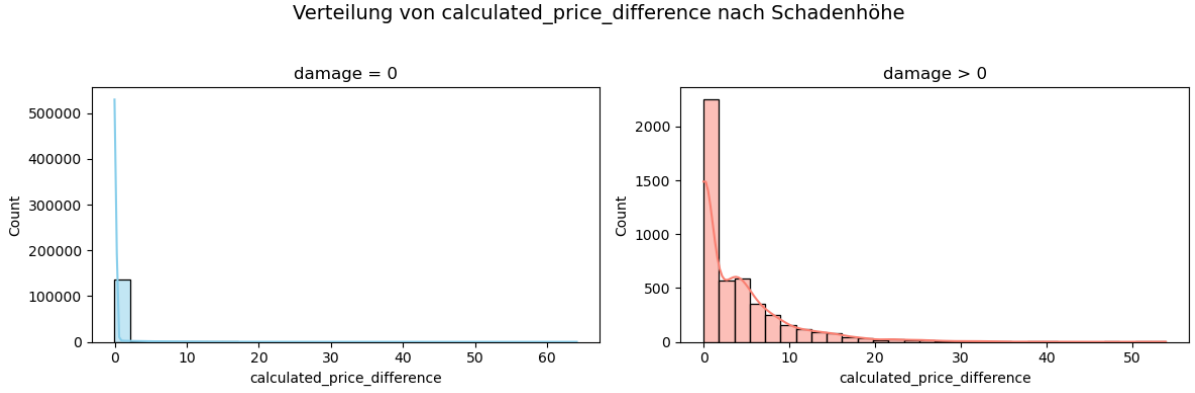
Transaktionen mit Schaden (damage > 0) zeigen tendenziell höhere Warenkorbsummen, eine höhere Anzahl gekaufter Artikel sowie eine längere Transaktionsdauer als korrekte Transaktionen. Alle drei Merkmal sind erwartungsgemäß stark miteinander korreliert. Dies kann dadurch erklärt werden, dass bei zunehmendem Warenkorb die Wahrscheinlichkeit für Fehltransaktionen naturgemäß steigt (z.B. falsches Scannen oder versehentlich weggelassene Artikel). Es lässt sich schließen, dass größere Einkäufe ein höheres Verlustrisiko bergen.



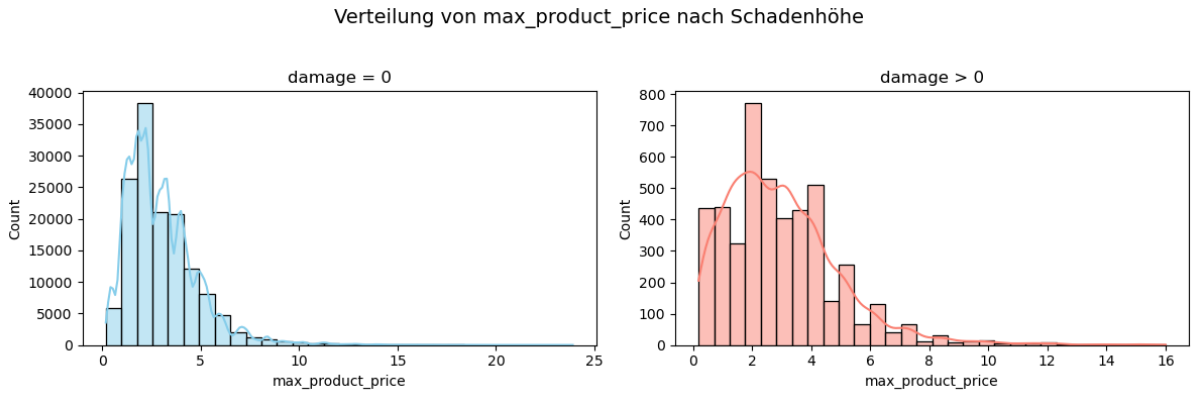




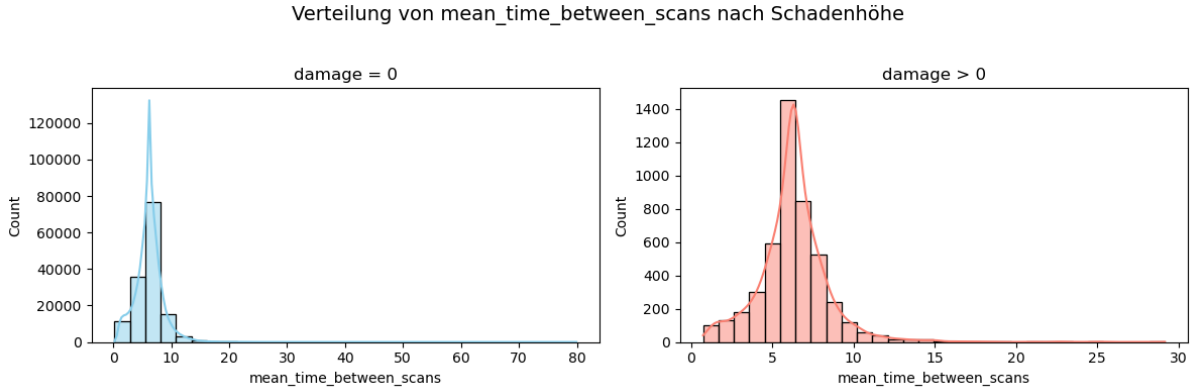
Transaktionen mit Schaden zeigen deutlich häufiger hohe Werte bei der calculated_price_difference, d.h. der Differenz zwischen der Summe der einzelnen Artikelpreise und der ausgewiesenen Gesamtsumme. Dies spricht für inkonsistente Preise oder fehlerhafte Scans als mögliche Verlustursache.



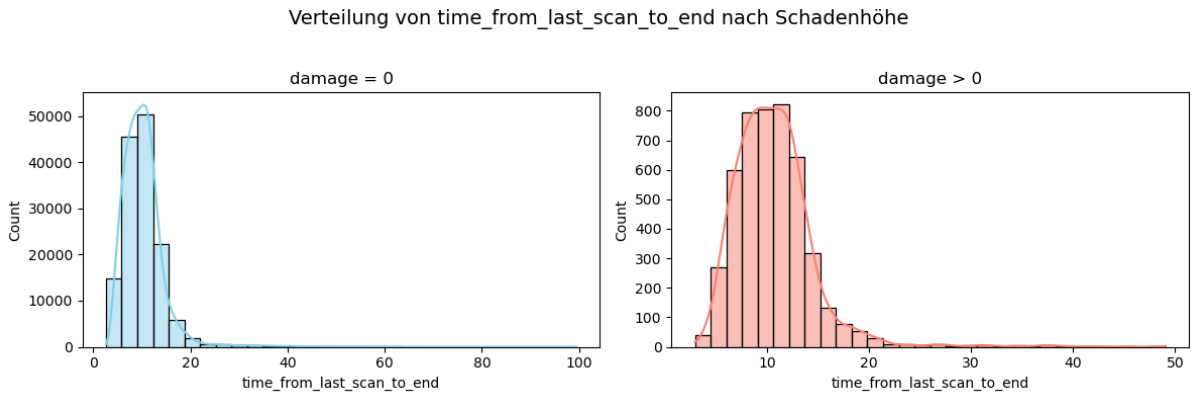
Einzelne besonders teure Produkte (max_product_price) treten bei Schadensfällen häufiger auf.



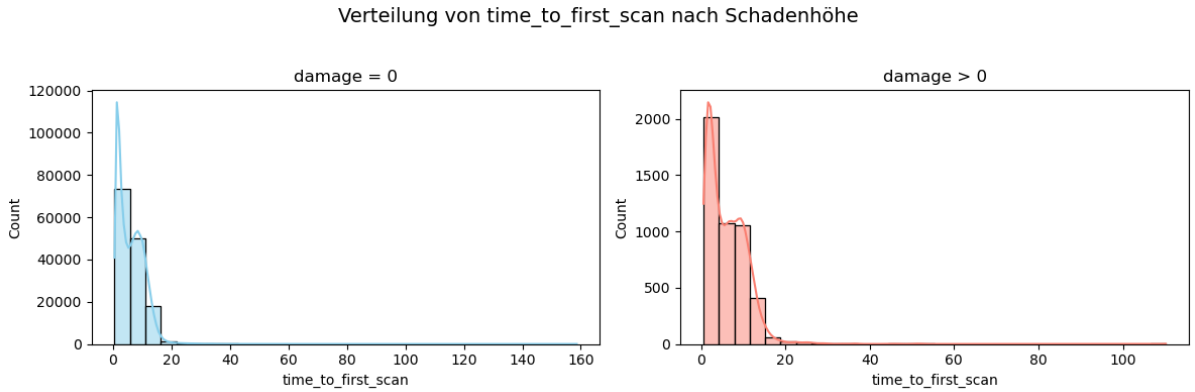
Das Scanverhalten ist bei Schadensfällen unregelmäßiger. Die mittlere Zeit zwischen zwei Scanvorgängen zeigt eine breitere Streuung bei Transaktionen mit Verlusten.



Die Zeit vom letzten Scan bis zum Bezahlabschluss ist bei schadensbehafteten Transaktionen variabler und potenziell länger. Dies könnte auf gezieltes Verzögern oder Unsicherheit hinweisen.



Schadensfälle benötigen im Durchschnitt etwas länger bis zum ersten Scan, was auf Unsicherheit, Ablenkung oder Vorbereitung hindeuten könnte.



Um Extremwerte in den numerischen Variablen zu identifizieren, wurde für jedes Feature der Z-Score berechnet und gezählt, wie viele Beobachtungen einen absoluten Z-Score über 3 aufweisen (entspricht grob einer Abweichung > 3 Standardabweichungen vom Mittelwert).

Diese Analyse erlaubt Rückschlüsse auf mögliche Fehleingaben, Sondereffekte oder systematisch auffällige Teilgruppen in den Daten.

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

Die Vermutung liegt nahe, dass die beobachteten Extremwerte nicht als Störgröße, sondern als relevante Erklärkraft interpretiert werden können. Eine detaillierte Analyse findet sich in den folgenden Abschnitten.

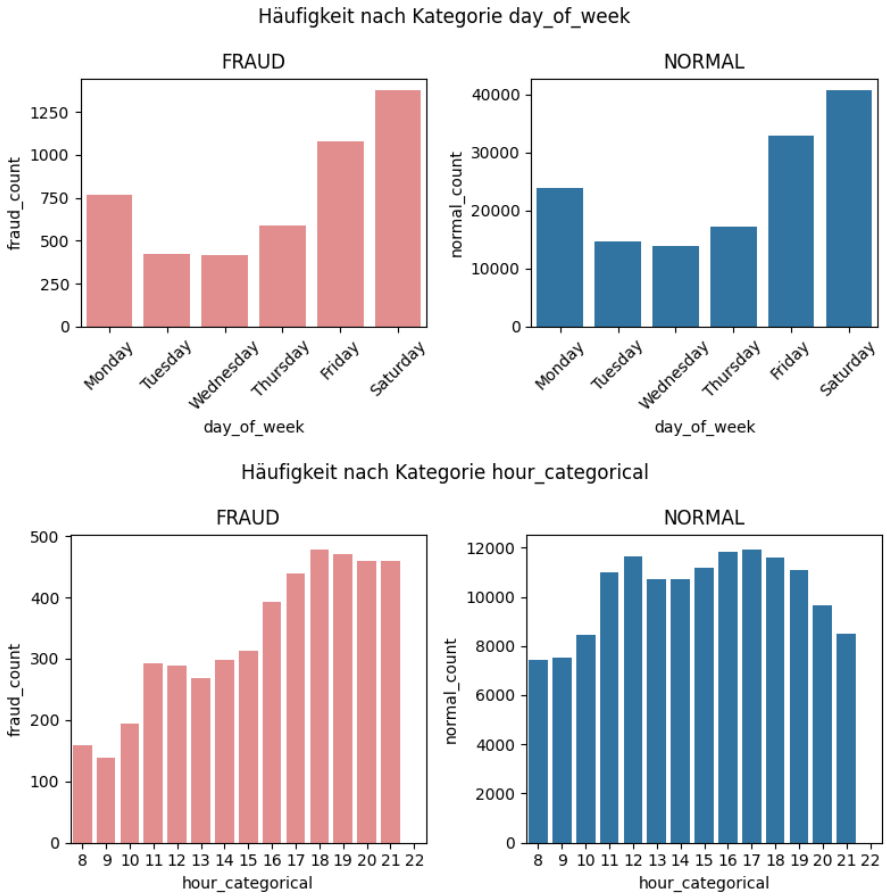
2. Analyse kategorialer Attribute

Verteilung der Kategorien

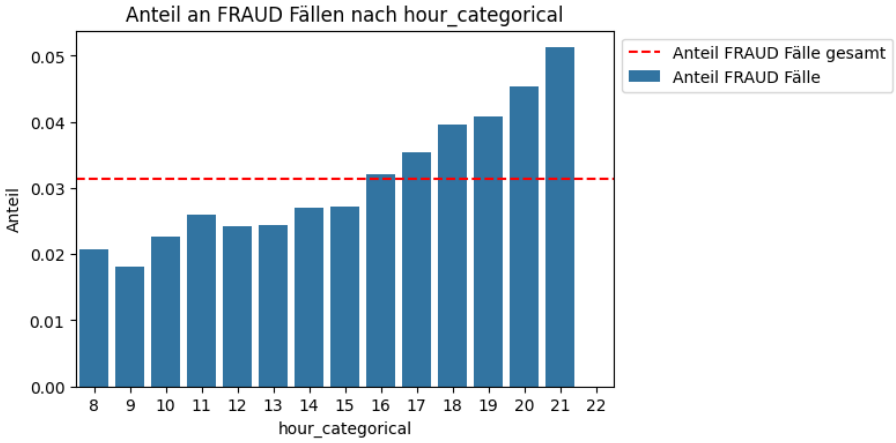
Die Häufigkeitsverteilungen aller kategorialen Variablen wurden mittels Balkendiagrammen visualisiert. Dabei zeigten sich teils starke Unausgeglichenheiten zwischen den Klassen.

Der Großteil der Transaktionen findet zwischen 12 und 19 Uhr statt, mit besonders hoher Frequenz an Samstagen und Feiertagen. Werktage wie Dienstag und Mittwoch weisen hingegen das geringste Transaktionsvolumen auf.

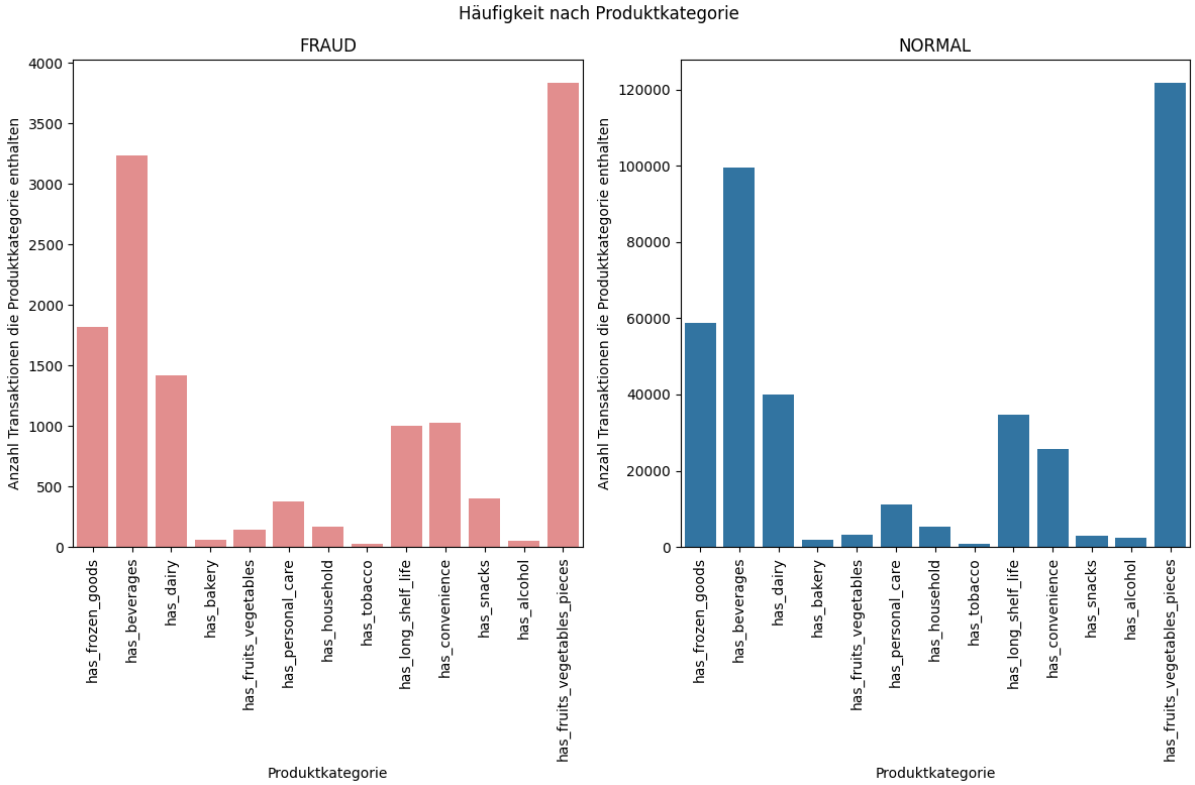
Während der Wochentag insgesamt keinen nennenswerten Einfluss auf den Anteil der *FRAUD*-Fälle hat, zeigt sich über den Tagesverlauf hinweg ein deutliches Muster: Die Anzahl der *FRAUD*-Fälle steigt kontinuierlich bis etwa 18 Uhr an und verbleibt anschließend auf einem erhöhten Niveau.



Betrachtet man den Anteil der *FRAUD*-Fälle relativ zur Gesamtzahl der Transaktionen, so ist dieser insbesondere in den späten Abendstunden überproportional hoch.

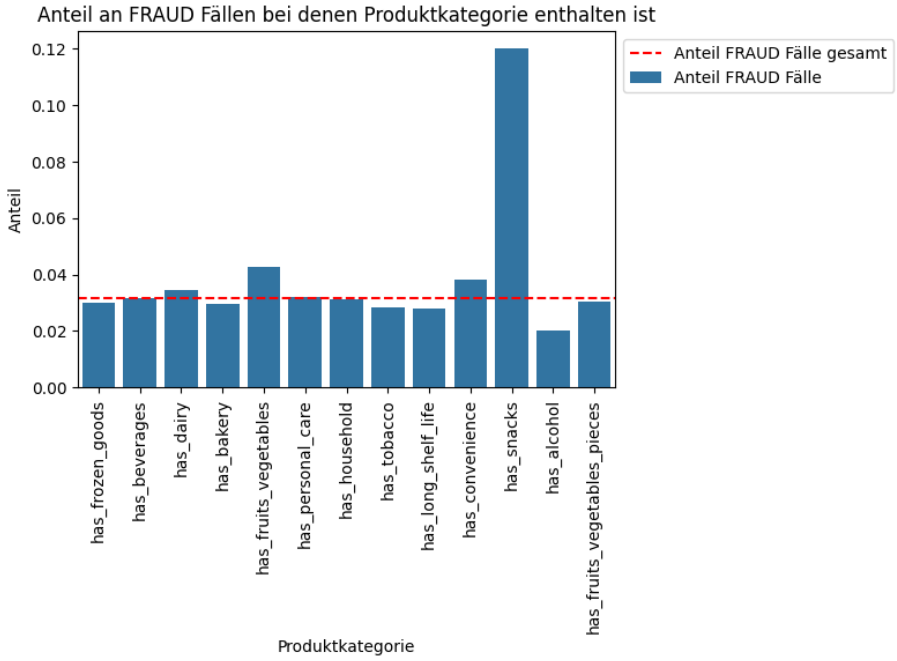


Einige Produktkategorien wie Früchte/Gemüse, Getränke oder Tiefkühlwaren sind stark vertreten. Andere wie Backwaren, Tabak oder Snacks kommen sehr selten vor. Generell lässt sich feststellen: Produktkategorien, die in vielen Transaktionen vertreten sind, treten auch häufiger in als *FRAUD* gelabelten Transaktionen auf – und umgekehrt.

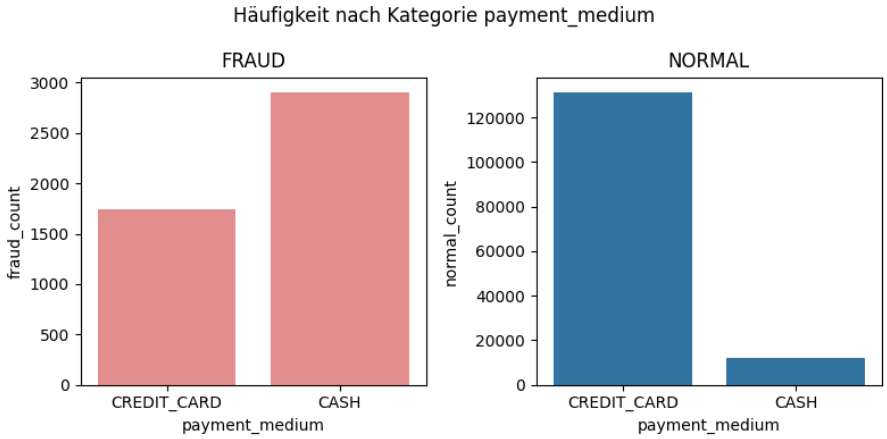


Wie die folgende Grafik zeigt, gibt es jedoch Ausnahmen von dem generellen Muster. So sind Transaktionen, die *Snacks* enthalten, unter den *FRAUD*-Fällen deutlich überrepräsentiert. Auch *Früchte und Gemüse* (nach Gewicht verkauft) sowie *Fertiggerichte* treten in *FRAUD*-Transaktionen etwas häufiger auf als im Gesamtdatensatz.

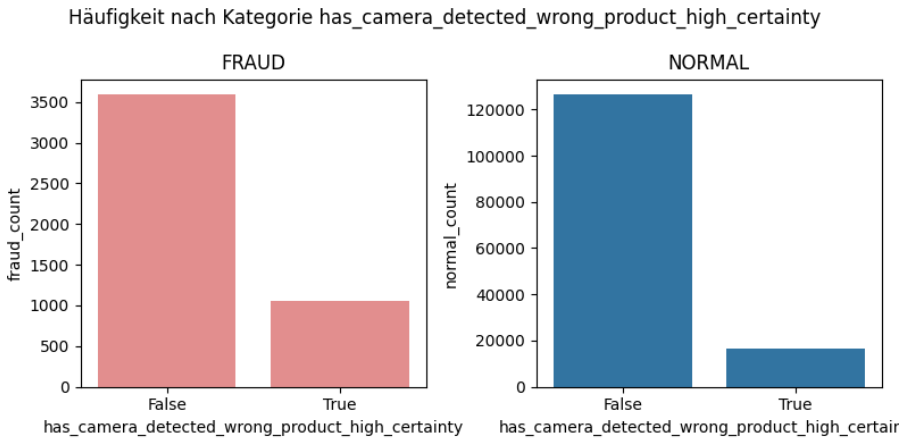
Auf der anderen Seite sind insbesondere *alkoholische Getränke*, aber auch *Tabakwaren* und *haltbare Produkte* in *FRAUD*-Fällen unterdurchschnittlich vertreten. Im Fall von *Alkohol* und *Tabak* könnten die verpflichtenden Alterskontrollen durch das Personal als abschreckender Faktor wirken und so das Risiko von FRAUD-Fällen reduzieren.



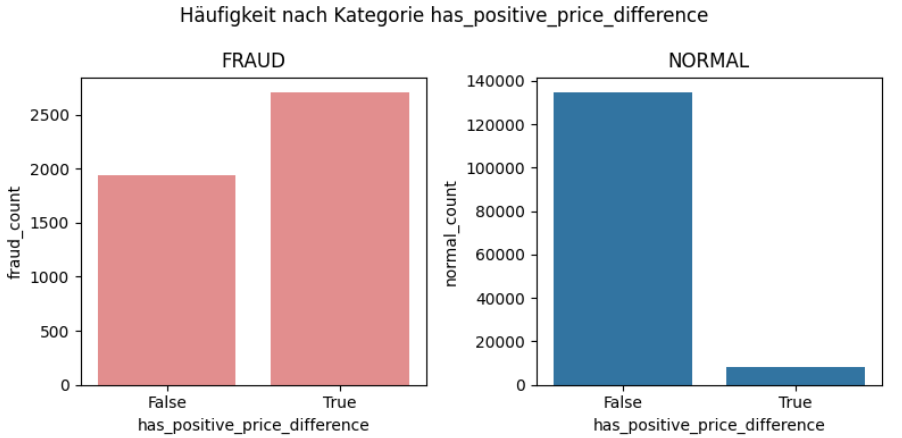
Die auffälligste Differenz zwischen *FRAUD*- und *NORMAL*-Transaktionen besteht im gewählten Zahlungsmittel: Während im Gesamtdatensatz überwiegend mit Kreditkarte gezahlt wird und die Barzahlung eher selten ist, erfolgt die Bezahlung in *FRAUD*-Fällen fast doppelt so häufig in bar. Es liegt nahe, dass Bargeld bevorzugt wird, um potenzielle Rückverfolgung zu vermeiden.



Das Vorhandensein eines von der Kamera nicht als ähnlich erkannten Produkts innerhalb einer Transaktion (hier nur Fälle in denen *camera_certainty* > 0.8) stellt zwar keinen eindeutigen Hinweis auf einen *FRAUD*-Fall dar, tritt jedoch in solchen Fällen erwartungsgemäß signifikant häufiger auf als in regulären Transaktionen.



Fälle, in denen der errechnete Nominalpreis nicht mit dem tatsächlich gezahlten Betrag übereinstimmt, machen über die Hälfte aller *FRAUD*-Fälle aus. Die unrechtmäßige Inanspruchnahme von Rabatten stellt somit vermutlich die häufigste Betrugsmasche dar.



Chi^2-Test zur Zielvariablen „FRAUD“

Einige Variablen wie has_feedback, feedback_categorical, day_of_week, location und payment_medium zeigen hochsignifikante Abweichungen zwischen FRAUD/NORMAL. Diese Variablen sind potenziell nützlich für Klassifikationsmodelle.

feature	chi2	p_value	significance
has_positive_price_difference	17833.647	0	*** sehr signifikant
payment_medium	14853.762	0	*** sehr signifikant
has_snacks	886.90327	6.90E-195	*** sehr signifikant
has_camera_detected_wrong_product_high_certainty	526.93111	1.31E-116	*** sehr signifikant
has_missing	461.54677	2.21E-102	*** sehr signifikant
hour_categorical	405.43215	9.09E-78	*** sehr signifikant
daytime	368.5321	1.45E-79	*** sehr signifikant
has_camera_detected_wrong_product	222.31929	2.82E-50	*** sehr signifikant
has_voided	55.78841	8.07E-14	*** sehr signifikant
feedback_categorical	54.2629	4.64E-11	*** sehr signifikant
has_convenience	49.314674	2.18E-12	*** sehr signifikant
feedback_high	25.528508	4.36E-07	*** sehr signifikant
has_fruits_vegetables_pieces	23.219589	1.45E-06	*** sehr signifikant
has_long_shelf_life	17.761352	2.50E-05	*** sehr signifikant
feedback_top	16.467099	4.95E-05	*** sehr signifikant
has_dairy	16.006413	6.31E-05	*** sehr signifikant
has_sold_by_weight	13.426353	0.0002481	*** sehr signifikant
has_fruits_vegetables	13.426353	0.0002481	*** sehr signifikant
day_of_week	13.193464	0.0216317	** signifikant
has_alcohol	9.722942	0.0018198	*** sehr signifikant
has_age_restricted	8.508482	0.0035349	*** sehr signifikant
has_frozen_goods	7.1599793	0.0074548	*** sehr signifikant
month	7.0422147	0.7956546	n.s.
feedback_middle	6.9276351	0.0084874	*** sehr signifikant
feedback_low	4.0463843	0.0442661	** signifikant
store_id	2.5431533	0.6369256	n.s.
location	2.5431533	0.6369256	n.s.
urbanization	2.3092761	0.3151716	n.s.
cash_desk	1.3846981	0.7091255	n.s.
has_feedback	0.6303315	0.4272338	n.s.
has_tobacco	0.1785727	0.6726027	n.s.
has_personal_care	0.1728052	0.6776307	n.s.
has_bakery	0.1555697	0.6932686	n.s.
has_beverages	0.041883	0.837843	n.s.
has_household	0.0056271	0.9402038	n.s.

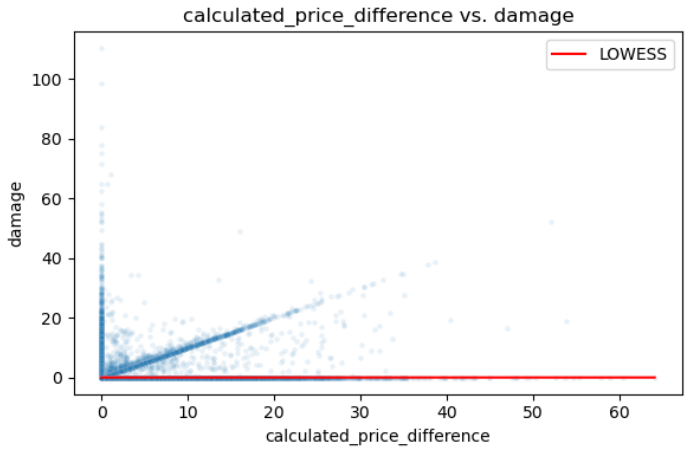
3. Nichtlineare Zusammenhänge zwischen Attributen und Schadenshöhe

Zur Analyse potenzieller nichtlinearer Zusammenhänge zwischen numerischen Features und der Zielgröße damage wurden zwei methodische Ansätze kombiniert:

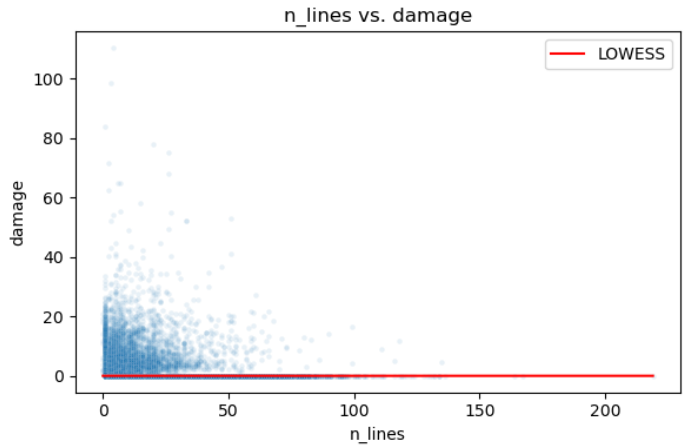
1. **LOWESS-Glättung** (lokal gewichtete Regressionslinien) zur visuellen Trendanalyse
2. **Korrelationen nach Pearson und Spearman** zur quantitativen Bewertung linearer und monotoner Zusammenhänge

Die Mehrheit der Attribute zeigt keine signifikanten Trends im Sinne eines Zusammenhangs mit der Schadenhöhe. Die Glättungslinien verlaufen in fast allen Fällen nahezu horizontal, was auf eine geringe Erklärkraft der Einzelvariablen hinweist.

Eine Ausnahme bildet das Feature calculated_price_difference (Preisdifferenz zwischen der ausgewiesenen Gesamtsumme und der Summe der einzelnen Produkte im Warenkorb), bei dem im mittleren Wertebereich eine leichte Zunahme der Schadenhöhe sichtbar wird. Hier zeigt sich auch im Scatterplot eine höhere Streuung, was auf eine potenziell komplexere Beziehung hinweist. Die Vermutung liegt nahe, dass ein Wert von diesem Attribut größer Null bereits sehr gut auf Schäden hindeutet, dass die Schadenhöhe allerdings nicht gut linear aus diesem Wert abgeleitet werden kann.



Auch die Anzahl der gekauften Artikel und der Einkaufsbetrag sind erwähnenswert: Sie zeigen eine starke Konzentration vieler Fälle im unteren Bereich, aber keine klaren Muster hinsichtlich zunehmender Schadenhöhe.



Alle anderen Attribute (z. B. Transaktionszeit des gesamten Einkaufs, Scan-Zeiten der einzelnen Artikel, Tage seit Einführung des Kamerasystems) zeigen keine systematische Veränderung der Zielvariable entlang des Attribut-Werts.

Korrelationsanalyse

Die Spearman-Korrelationen mit der Zielgröße damage bestätigen diese Beobachtung: Die höchsten Werte finden sich bei:

- calculated_price_difference (~ 0.09)
- total_amount (~ 0.06)
- n_lines (~ 0.05)

Alle übrigen Attribute liegen unterhalb der Schwelle von 0.05 und gelten damit als vernachlässigbar in Bezug auf monotone Zusammenhänge.

Schlussfolgerung

Einzelne numerische Attribute erklären nur einen sehr geringen Teil der Varianz von damage. Eine nichtlineare Modellierung auf Basis dieser Einzelmerkmale erscheint wenig erfolgversprechend. Kombinierte, multivariate Modelle mit Interaktionstermen sind zur

Abbildung relevanter Zusammenhänge wesentlich besser geeignet (vgl. Abschnitt: Regressionsmodellierung).

4. Regressionsmodellierung

Ein zentrales Ziel der explorativen Analyse war es, die erklärenden Variablen (Features) hinsichtlich ihrer statistischen und praktischen Relevanz für die Zielgrößen 'damage' (numerisch) und 'label' (binär: 'FRAUD' vs. 'NORMAL') zu untersuchen. Dabei wurde ein zweistufiges Vorgehen angewendet, univariate Signifikanztests pro Feature bzw. multivariate Modellbildung mit schrittweiser Reduktion.

Univariate Regressionsanalyse

Für jedes Feature wurde zunächst ein einfaches Regressionsmodell berechnet, das das jeweilige Merkmal als einzigen Prädiktor enthielt. Je nach Datentyp des Features wurde folgendes Vorgehen gewählt:

- Kategoriale Variablen: Chi^2-Test zur Messung der Abhängigkeit vom Ziel.

- Numerische Variablen: Regressionsmodell mit p-Wert und erklärter Varianz (Pseudo-R^2 bei Klassifikation, R^2 bei Regression).

Die für die Klassifikation interessanten Variablen sind:

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

Und die für die Vorhersage der Schadenshöhe:

feature	significance	relevance
payment_medium	sehr signifikant	sehr relevant
hour	sehr signifikant	weniger relevant
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

Die vollständigen Ergebnisse sind in den Dateien 'feature_analysis_damage.xlsx' und 'feature_analysis_label.xlsx' dokumentiert.

Multivariate Regressionsanalyse

Für beide Zielgrößen wurde jeweils ein multivariates Regressionsmodell mit allen Features (inkl. Interaktionsterme und polynomiellen Komponenten) erstellt und durch schrittweise Elimination

vereinfacht. Im fertigen Modell befinden sich eine Vielzahl von Attributen. Dies gilt es, im weiteren Verlauf des Projekts genauer zu untersuchen, um die Robustheit und Generalisierbarkeit des Modells sicher zu stellen. In der aktuellen Form des Modells ist die Prognosegüte bereits erstaunlich hoch. Dabei fand die Berechnung des Modells auf einer Zufallsauswahl aus den zur Verfügung stehenden Daten statt, die Evaluation auf dem restlichen Datensatz (80%/20%) statt.

```
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
```

Bei der Regression auf die Zielvariable „damage“ lag die Güte des Modells deutlich niedriger. Das lässt sich dadurch erklären, dass erstens der Schadensbetrag im Schadensfall eine breite Streuung hat und sich nicht gut aus den erklärenden Variablen präzise vorhersagen lässt, andererseits enorm viele Fälle mit einem Schaden von Null (kein Betrug) vorkommen. Dadurch wird das Modell verzerrt.

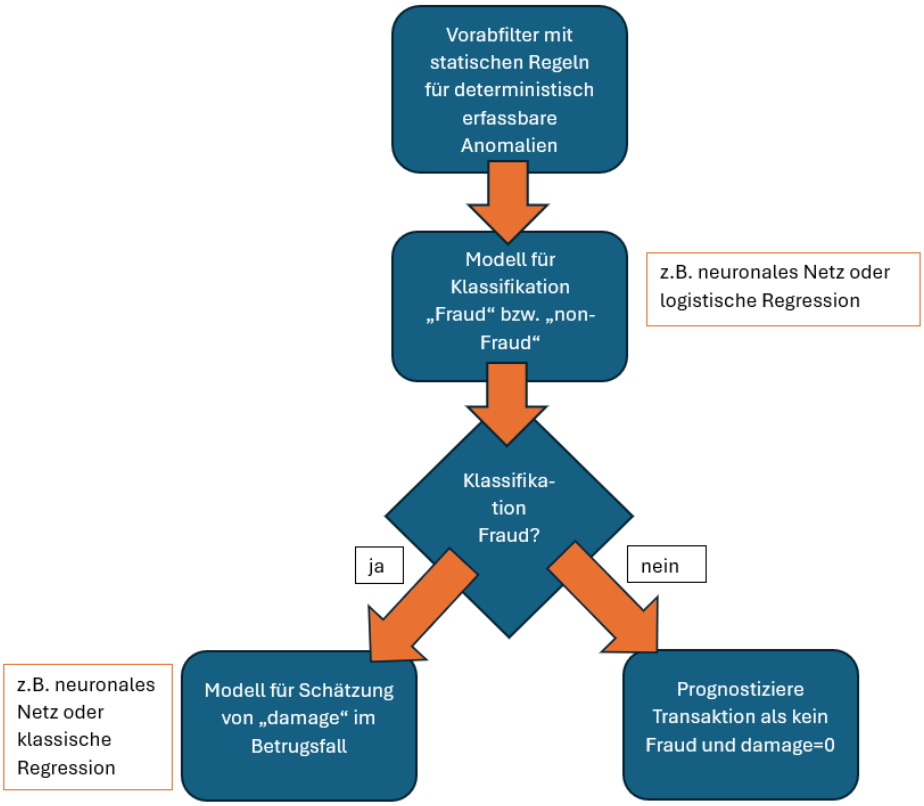
```
Damage-Modell:
 R² Test:  0.137
 R² Train: 0.136

 RMSE Test:  1.754
 RMSE Train: 1.721
```

Mit komplexeren Modellen mit Interaktionstermen und / oder polynomiellen Termen konnte zwar eine Verbesserung auf der Trainingsmenge, nicht jedoch auf der Testmenge erreicht werden. Dies spricht für Überanpassung des Modells an den Trainingsdatensatz und eine schlechte Generalisierung.

Nächste Schritte

Bis jetzt haben wir die komplexe Beziehung von „calculated_price_difference“ zu „damage“ noch nicht im Detail berücksichtigt. Auch haben wir nur einfachste Modelle gerechnet (z.B. keine neuronalen Netze) und die strukturierte Kürzung um redundante Attribute zum Zwecke der Generalisierbarkeit steht noch aus. Ein möglicher Ansatz für den nächsten Meilenstein ist nun, das Vorhersagemodell in drei Teile aufzuspalten. Ein erster statischer Test auf Inkonsistenzen, der verdächtige Transaktionen benennet. Auf diesen Regeln aufbauend ein zweites Modell, das lediglich eine Betrugsklassifikation vornimmt, d.h. eine Prognose, ob die Transaktion fehlerhaft

ist oder nicht. Erst im dritten Schritt wird dann in einem weiteren Modell die Schadenshöhe modelliert. Es ist anzunehmen, dass das letzte Modell deutlich besser funktioniert, wenn ein Schaden von Null (also kein Betrug laut Klassifikationsmodell) bereits vorab herausgefiltert wird und nur noch ein tatsächlicher Schaden im Schadensfall vorhergesagt werden muss. Durch die gut Vorhersagbarkeit von falschen Transaktionen, kann zudem bereits eine Kontrolle ausgelöst werden und so der Schaden aktiv verhindert werden, unabhängig von deren Höhe. Die skizzierte Architektur ist in folgendem Schaubild verdeutlicht:



Zusammenfassung und Ausblick

Die bereitgestellten Daten wurden erfolgreich integriert, bereinigt und analysiert. Es liegen valide und repräsentative Merkmalsräume für die Modellbildung vor. Die nächste Phase – **Modellentwicklung** – kann auf einer stabilen, sauberen Datenbasis aufsetzen.

Die Bewertung der Modelle erfolgt in enger Abstimmung mit der Wertkauf GmbH auf Basis einer **Kostenfunktion**, die Fehlklassifikationen nach ihrem finanziellen Einfluss unterschiedlich gewichtet.