Dokumentation – Meilenstein 2: Datenaufbereitung und Datenanalyse

Vorabanalyse und Vorbemerkungen

Der zweite Meilenstein im Rahmen des DASC-PM umfasst die **Datenbereitstellung**, **Datenprüfung** 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.

**Verwendung gelabelter Daten auf Basis statistischer Repräsentativität**

Für sämtliche Analysen und Modellierungen im Rahmen dieses Projekts werden ausschließlich die **gelabelten Transaktionen** berücksichtigt (mit label = FRAUD oder label = NORMAL).  
Transaktionen mit label = UNKNOWN wurden ausgeschlossen.  
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 durch den **Lernprozess des Kamerasystems** bedingt sein.

**Differenzen zwischen tatsächlichem und rechnerisch korrektem sales\_price**

Bei der rechnerischen Überprüfung des in den Transaktionszeilen angegebenen sales\_price anhand des Produkts aus Menge und Preis pro Einheit (basierend auf products.csv) wurden in ca. **5.400 Einzelfällen** signifikante Abweichungen festgestellt – ausschließlich rund **30 %** bezogen auf den rechnerisch erwarteten Wert.  
Nach Rücksprache mit der Wertkauf GmbH könnte dies auf unrechtmäßig in Anspruch genommene **Standardrabatte für Ware kurz vor Ablauf des Mindesthaltbarkeitsdatums** zurückzuführen sein.  
Diese Fälle wurden in der nachfolgenden Analyse möglicherweise noch nicht vollständig berücksichtigt.

Nach derzeitigem Stand betreffen diese Abweichungen ca. **2.713 Transaktionen** mit rund **5.400 Einzelpositionen** und verursachen einen **Gesamtschaden von ca. 18.800 €**, was etwa **53 % der gesamten erfassten Schadenssumme** entspricht.  
Weitere Analysen sowie Vorschläge zur Reduzierung dieser Verluste werden im weiteren Projektverlauf erarbeitet.

**Analyse des Storno-Prozesses**

Die Analyse des Stornierungsprozesses führte zu mehreren Auffälligkeiten und erforderte Rücksprachen mit der Wertkauf GmbH.  
Die daraus resultierenden Erkenntnisse bedürfen noch einer detaillierten Analyse und Bewertung und werden – sofern relevant – in zukünftige Modellversionen integriert.

Darüber hinaus wurden rund **500 Positionen** identifiziert, bei denen der sales\_price genau **0,00 €** beträgt.  
Diese Fälle betreffen ausschließlich **Stornierungen mit dem Label FRAUD**, was auf mögliche Prozesslücken hinweist.

**Berechnete Merkmale**

Auf Basis der vorhandenen Spalten wurden zusätzliche Attribute generiert, darunter:

* Transaktionsdauer
* Uhrzeit des Einkaufs
* Wochentag des Einkaufs
* Monat des Einkaufs

Insbesondere die **Uhrzeit** zeigte sich als signifikant für die Unterscheidung zwischen FRAUD und NORMAL.

- David: Ich erstelle hier noch zwei Abschnitte zur Feedback-Spalte und dem Kamerasystem.

- Matthias: M.E. neigen ehrliche Kunden eher zur Abgabe einer schlechten Bewertung. Falls das so ist: müssten wir das noch berücksichtigen?

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
* **Join-Strategien:** Zur Kombination von Transaktionen und Zeileninformationen wurde über die Spalte `transaction\_id` ein Join durchgeführt. Weitere Joins mit Produktdaten (`product\_id`) und Store-Daten (`store\_id`) wurden durchgeführt.

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.

1. **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.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Transaktionen mit Schaden (damage > 0) zeigen tendenziell höhere Warenkorbsummen und eine höhere Anzahl gekaufter Artikel als korrekte Transaktionen. 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.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.

KI-generierte Inhalte können fehlerhaft sein.

Die Transaktionen mit Schaden dauern im Mittel länger. Mögliche Ursachen sind Unsicherheiten, Störungen oder bewusste Manipulationsversuche.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

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.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Einzelne besonders teure Produkte (max\_product\_price) treten bei Schadensfällen häufiger auf.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

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

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

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

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

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

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

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.

## Ein Bild, das Text, Screenshot, Schrift, Zahl enthält. KI-generierte Inhalte können fehlerhaft sein.

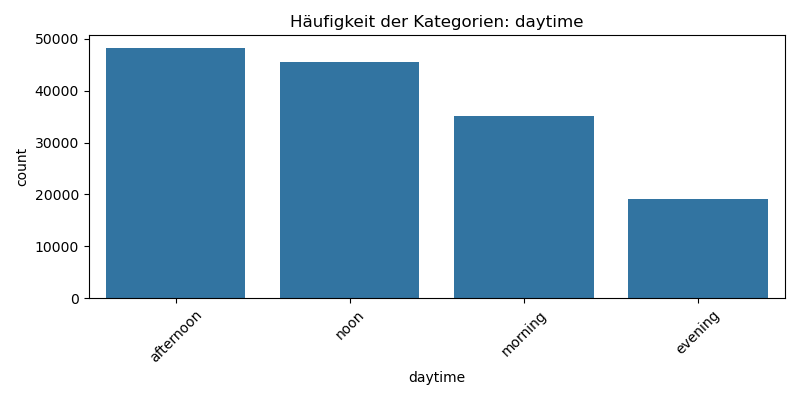
Die Vermutung liegt nahe, dass die beobachteten Extremwerte nicht als Störgröße, sondern als relevante Erklärkraft interpretiert werden können. Eine detaillierte Analyse findet sich in den folgenden Abschnitten.

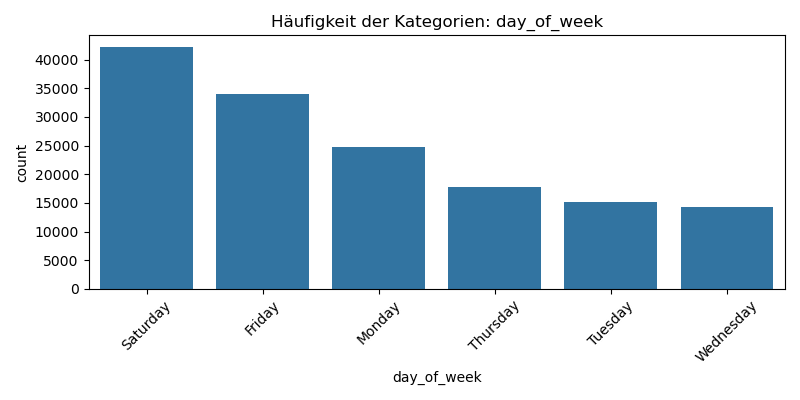
1. **Analyse kategorialer Attribute**

**Verteilung der Kategorien**

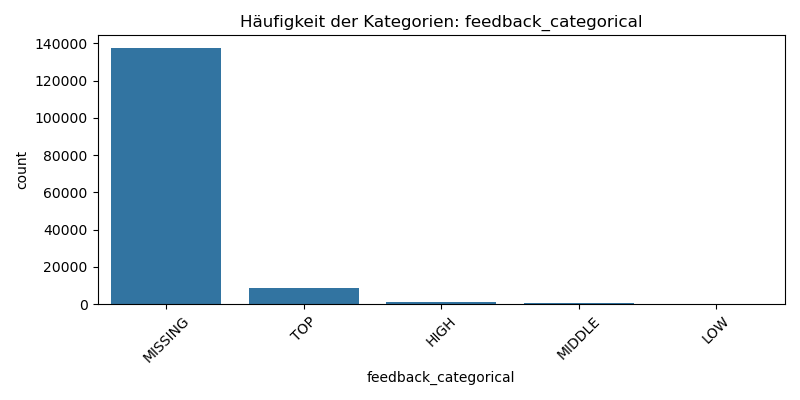
Die Häufigkeitsverteilungen aller kategorialen Variablen wurden mittels Balkendiagrammen visualisiert. Dabei zeigten sich teils starke Unausgeglichenheiten zwischen den Klassen.

Transaktionen finden überwiegend nachmittags statt, besonders an Samstagen. Werktage wie Mittwoch und Dienstag sind am wenigsten frequentiert.





Der überwiegende Teil der Transaktionen enthält kein Feedback (Kundenbefragung zur Zufriedenheit). Innerhalb der Feedback-Kategorien dominiert daher die Klasse "MISSING", andere wie "LOW" und "MIDDLE" kommen nahezu nie vor.



Einige Produktkategorien wie Getränke oder Tiefkühlwaren sind stark vertreten. Andere wie Backwaren, Tabak oder Snacks kommen sehr selten vor.

Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

KI-generierte Inhalte können fehlerhaft sein.Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

KI-generierte Inhalte können fehlerhaft sein.Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

KI-generierte Inhalte können fehlerhaft sein.

Der Großteil der Einkäufe erfolgt per Kreditkarte, Barzahlungen sind deutlich seltener.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Chi²-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.

Ein Bild, das Text, Screenshot, Karte Menü, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

1. **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.

Ein Bild, das Text, Screenshot, Reihe, Diagramm enthält.

KI-generierte Inhalte können fehlerhaft sein.

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.

Ein Bild, das Text, Screenshot, Software, Display enthält.

KI-generierte Inhalte können fehlerhaft sein.

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).

1. **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: Chi²-Test zur Messung der Abhängigkeit vom Ziel.

- Numerische Variablen: Regressionsmodell mit p-Wert und erklärter Varianz (Pseudo-R² bei Klassifikation, R² bei Regression).

Die für die Klassifikation interessanten Variablen sind:

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Und die für die Vorhersage der Schadenshöhe:

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

Die vollständigen Ergebnisse sind in den Dateien 'feature\_analysis\_damage.xlsx' und 'feature\_analysis\_label.xlsx' dokumentiert.

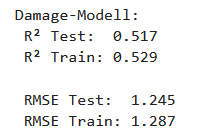
**Multivariate Regressionsanalyse**

Für beide Zielgrößen wurde jeweils ein multivariates Regressionsmodell mit allen Features (inkl. Interaktionsterme und polynomiellen Komponenten) erstellt und durch schrittweise Elimination vereinfacht. Dabei wurden nur Variablen mit signifikanter Erklärkraft im Modell behalten. Die Berechnung des Modells fand auf einer Zufallsauswahl aus den zur Verfügung stehenden Daten statt, die Evaluation auf dem restlichen Datensatz (80%/20%). Ein einfaches Modell ohne Interaktionen oder polynomielle Terme erreicht bereits eine erstaunlich gute Prognosegüte bei der Klassifikation:

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

Bei der Regression auf die Zielvariable „damage“ lag die Güte des Modells niedriger, allerdings ebenfalls überraschend gut für einen ersten Ansatz. Über 50% der Varianz der Zielvariable „damage“ können erklärt werden, sowohl auf dem Trainings- als auch auf dem Testdatensatz:



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“ berücksichtigt. Auch haben wir nur einfachste Modelle gerechnet. Ein möglicher Ansatz für den nächsten Meilenstein ist nun, das Vorhersagemodell in drei Teile aufzuspalten:

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

KI-generierte Inhalte können fehlerhaft sein.

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 Modelle werden anhand eines gewichteten Kostenmodells (Bewertungsfunktion) in enger Absprache mit der Wertkauf GmbH bewertet, das Fehlklassifikationen monetär gewichtet.

AB HIER:

(kann weg, wenn keine Änderung mehr notwendig: Modell ist oben als Screenshot eingefügt, damit sie sich einfacher verschieben und vergrößern / verkleinern lässt

Vorabfilter mit statischen Regeln für deterministisch erfassbare Anomalien

Modell für Klassifikation „Fraud“ bzw. „non-Fraud“

z.B. neuronales Netz oder logistische Regression

z.B. neuronales Netz oder klassische Regression

nein

ja

Prognostiziere Transaktion als kein Fraud und damage=0

Modell für Schätzung von „damage“ im Betrugsfall

Klassifika-tion Fraud?