Data Audit Report – Wertkauf GmbH

1. Datenherkunft & Annahmen

Die bereitgestellten Transaktionsdaten wurden nach unserer Kenntnis gemäß der Kassensicherungsverordnung (KassenSichV) erfasst und durch zertifizierte technische Sicherheitseinrichtungen (TSE) abgesichert. Weitere technische Dokumentationen oder Auditberichte liegen nicht vor. Die Daten enthalten:

- Gelabelte Transaktionen (FRAUD / NORMAL)

- Produkt- und Artikeldaten

- Standortinformationen (Filiale, Urbanisierung)

- Zeitstempel und Kamerasystemdaten

Hinweis auf systematische Lücken: Negative Schadensfälle (z. B. vergessene Ware im Geschäft) fehlen vollständig. Dies führt zu einer strukturellen Verzerrung in der Zielgröße 'damage'.

2. Datenstruktur & Quellen

Die Datenmenge besteht aus zwei CSV-Dateien mit Stammdaten und vier Parquet-Dateien mit Transaktions- und Positionsdaten. Die Datenbasis umfasst rund 1,48 Mio. Transaktionen, von denen rund 148.000 einer manuellen Kontrolle unterzogen wurden.

Strukturelle Merkmale:

- Transaktionen als Hauptobjekte (Preis, Artikel etc.)

- Zahlreiche zusätzliche Attribute (benötigte Scanzeit, Tageszeit des Einkaufs, Indikationen des Kamerasystems etc.), teils kategorial, teils numerisch

Die Trainingsdaten stammen aus dem Jahr 2022-2023, die Testdaten aus 2024. Eventuelle Trends in den Daten bzw. Verschiebungen zwischen Trainings- und Testdaten sind zu beachten und wurden nicht im Detail untersucht.

3. Datenqualität & fehlende Werte

Eine detaillierte Prüfung ergab folgende Beobachtungen:

- Fehlende Werte hauptsächlich bei 'customer_feedback', 'weight', 'valid_to', 'camera_certainty', 'camera_product_similar'

- Systematische Qualitätsunterschiede bei der Kameraauswertung zwischen Frühphase (vor Juli 2022) und der darauffolgenden Zeit

- Rundungsprobleme bei Preisen

- Teils inkonsistente Anzahl von Positionsdatensätzen zu n_lines

Fehlende Werte wurden – sofern sinnvoll – im Zuge der Datentransformation durch Imputation ersetzt (z. B. Modus oder Gruppenmittelwert) bzw. gelöscht. Komplexe oder unklare Felder wie 'camera_product_similar' wurden binär interpretiert oder exkludiert.

4. Datenkonsistenz & technische Prüfung

Die meisten Konsistenzprüfungen verliefen erfolgreich:

- Zeitstempel folgen plausibler Reihenfolge (Start < letzter Scan < Endzeitpunkt)

- Einzelpreis × Menge stimmt häufig, aber nicht immer mit Verkaufspreis überein

- Einige Transaktionen zeigen Differenzen zwischen Summe der Einzelpositionen und ausgewiesener Gesamtsumme

Diese Differenzen wurden über die Variable 'calculated_price_difference' quantifiziert und als Merkmal nutzbar gemacht.

5. Dokumentations- und Interpretationslücken

Im Datenmaterial fehlen offizielle technische Dokumentationen und ein vollständiges Data Dictionary. Dies führt zu Unsicherheiten bei der Interpretation einzelner Spalten.

Betroffen sind u. a.:

- feedback_categorical

- camera_product_similar & camera_certainty

- weight, valid_to

- urbane Standortmerkmale und Kassentypen

Empfehlung: Nachreichen einer variablenbezogenen Beschreibung zur langfristigen Nachvollziehbarkeit.

6. Kontrollstrategie & Labelvergabe

Die Zielvariablen 'label' und 'damage' wurden nicht automatisch erzeugt, sondern basieren auf stichprobenbasierten Kontrollen.

- Nur rund 148.000 von 1,48 Mio. Transaktionen wurden manuell kontrolliert

- FRAUD-Fälle machen davon rund 3 % aus

Die Wahrscheinlichkeit, dass eine Transaktion kontrolliert wird, hängt nicht direkt vom Inhalt der Transaktion ab, sondern unterliegt internen Prozessen.

Daraus ergibt sich ein Label Bias, der bei datengetriebener Analyse zu berücksichtigen ist.

7. Fazit & Empfehlung

Die vorliegende Datenbasis ist konsistent, strukturiert und grundsätzlich modellierungsfähig. Es bestehen allerdings Verzerrungen in der Labelvergabe und fehlende Dokumentationen.