Projekt: Verlustprävention an Selbstbedienungskassen

Data Audit Report zum Meilenstein 2 nach DASC-PM
DATENBEREITSTELLUNG

Projektkontext

Ziel des Projekts ist es, mit Hilfe von Transaktions- und Produktdaten Verluste (z. B. durch nicht korrekt erfasste Artikel) an Selbstbedienungskassen (SBK) zu identifizieren, zu modellieren und durch präventive Maßnahmen zu minimieren.

1. DATENBESCHREIBUNG

Die Wertkauf GmbH hat unserer Projektgruppe mehrere strukturierte Datensätze zur Verfügung gestellt, die sich auf Transaktionen an Selbstbedienungskassen beziehen. Die Daten beinhalten Informationen über vollständige Einkäufe, Kontrollklassifikationen sowie ergänzende Informationen zu Produkten und Filialen.

Die Datenquellen liegen in folgenden Dateien vor:

Datei	Inhalt	Relevanz
transactions_train.parquet	Metadaten zu Transaktionen inkl. Label (NORMAL, FRAUD, UNKNOWN) und Schaden (damage)	zentrale Grundlage für Klassifikation und Regressionsmodelle
transactions_lines_train.parquet	Detaillierte Zeilen zu gekauften Produkten pro Transaktion (z. B. Produkt, Preis, Gewicht, Voiding)	Grundlage für Feature-Engineering auf Artikel-Ebene
products.csv	Produktstammdaten inkl. Preis, Kategorie, Altersfreigabe etc.	Anreicherung der line-items
stores.csv	Standortinformationen (z. B. Urbanisierung, Bundesland, Öffnungsdatum)	mögliche erklärende Variable
transactions_test.parquet, transactions_lines_test.parquet	Entsprechend der Trainingsdaten, aber ohne Label für spätere Evaluation	wird im Modellvalidierungsschritt verwendet

- **transactions_train.parquet**

  Enthält 1.481.783 Transaktionen aus dem Trainingszeitraum, davon 148.025 kontrollierte Transaktionen (gelabelt mit NORMAL oder FRAUD), 4.656 davon mit erkanntem Betrug. Enthält u.a. Zeitstempel, Zahlungstyp, Kassennummer und Kundenfeedback.

- **transactions_lines_train.parquet**

  Enthält 15.793.671 einzelne Kassenzeilen (Produkte) zu den Transaktionen, inkl. Produkt-ID, Menge (Stück/Gewicht), Preis, Kamera-Sicherheitsklassifikation und Zeitstempel pro Scanvorgang.

- **products.csv**

  Enthält 8.120 Produkte mit Eigenschaften wie Kategorie, Preis, Gewicht, Beliebtheit, Altersfreigabe sowie Gültigkeitszeitraum.

- **stores.csv**

  18 Filialen mit Standortinformationen, Bundesland, Urbanisierungsgrad und Datum der Einführung von Selbstbedienungskassen (SBK).

2. METADATEN

Zusätzlich wurden von der Wertkauf GmbH Metadaten zur Verfügung gestellt, welche die fachliche Bedeutung der Spalten sowie Datenqualitätsanforderungen dokumentieren. Diese Informationen wurden bei der Datenprüfung und Vorbereitung berücksichtigt.

3. DATENQUALITÄT UND -BEREINIGUNG

- **Fehlende Werte:**
  - `customer_feedback`: Nur in ca. 105.000 von 1,48 Mio. Fällen vorhanden
  - `damage`: Nur bei kontrollierten Transaktionen vorhanden
  - `weight`: Teilweise fehlend bei Produkten, die nicht nach Gewicht verkauft werden
  - `valid_to`: Fehlend bei Produkten, die derzeit noch aktiv verkauft werden

- **Bereinigungsschritte:**
  - Konvertierung von Zeitspalten zu `datetime`
  - Behandlung von fehlenden Werten (Imputation bei Bedarf)
  - Entfernen oder Transformieren von Ausreißern in numerischen Spalten
  - Zusammenführung von Transaktionen und Produktinformationen für Feature Engineering

4. Datenaufbereitung

Durchgeführte Schritte

- **Zeitliche Merkmale extrahiert**:
  - „transaction_duration“ (in Sekunden) als Differenz aus transaction_end und transaction_start
  - „hour_of_day „aus transaction_start
  - „Month“ aus transaction_start
  -
- **Labels bereinigt**: Fokus nur auf kontrollierte Transaktionen (label ≠ UNKNOWN)
- **Join-Vorgänge durchgeführt**:
  - Transaktionen mit Transaktionszeilen (über id ↔ transaction_id)
  - Lines mit Produktdaten (über product_id)
  - Stores mit Transaktionen (über store_id)
- **Fehlende Werte identifiziert und ggf. imputiert**:
  - Z. B. customer_feedback oder weight (imputation geplant/nach Regelwerk)
- **Ausreißer identifiziert (z. B. via Z-Score oder IQR)**:
  - Für Variablen wie total_amount, transaction_duration
- **Mehrere signifikante Merkmale identifiziert**

- **Wesentliche Merkmale werden für die spätere Modellbildung in der Feature Map erfasst**

---

5. EXPLORATIVE DATENANALYSE (EDA)

Wichtiger Hinweis:
Die EDA wurde ausschließlich auf Basis der gelabelten Transaktionen (label ∈ {FRAUD, NORMAL}) durchgeführt. Der Grund: Nur für diese Daten lagen verlässliche Zielgrößen vor. Die Entscheidung beruht auf vorab durchgeführten Signifikanztests (t-Test und Chi-Quadrat-Test), die zeigten, dass diese Teilmenge **repräsentativ** für den Gesamtdatensatz ist.

Zur Vorbereitung der Modellierung wurde eine erste umfassende EDA durchgeführt:

Ziel: Strukturen, Auffälligkeiten und potenzielle Features erkennen

-Numerische Merkmale:

Statistische Kenngrößen (Mittelwert, Median, Standardabweichung, Ausreißer), Visualisierungen wie Histogramme und Boxplots. Außerdem: t-Tests zwischen gelabelten und ungelabelten Transaktionen zur Prüfung der Repräsentativität.

- Kategorische Merkmale

Häufigkeitstabellen, Chi-Quadrat-Tests, Visualisierung mit Balkendiagrammen und Heatmaps (z.B. Cramer's V) zur Bewertung von Zusammenhängen mit der Zielvariable label.

- Zeitvariablen:

Die Spalten `transaction_start` und `transaction_end` wurden genutzt, um eine neue Spalte `transaction_duration` (Transaktionsdauer in Sekunden) zu erstellen. Zusätzlich wurden die Tagesstunde und der Monat aus `transaction_start` extrahiert, um potenzielle zeitliche Muster zu erkennen.

Ergebnisse der Analyse (Visualisierungen & Grafiken):

Die Ergebnisse und Visualisierungen der explorativen Datenanalyse (z. B. Verteilungen, Korrelationen, Heatmaps, etc.) werden in einer separaten Präsentation ausführlich erläutert und bereitgestellt.

6. 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.

- Hinweis: Die vollständige technische Dokumentation inklusive aller verwendeten Skripte, Transformationsschritte und Explorationsgrafiken wird separat zur Verfügung gestellt.
- Strukturierung: Die Daten wurden in DataFrames organisiert, jeweils für Transaktionen, Produktzeilen, Produkte und Stores.
- Transaktionen mit Label ≠ UNKNOWN wurden gefiltert und separat gespeichert (`labeled`).

---

7. Ausblick (für Meilenstein 3: Modellierung)

- Klassifikationsmodell (Ziel: label) geplant
    - z. B. Random Forest, Gradient Boosting, Logistische Regression
- Alternativ/ergänzend: Regressionsmodell auf damage (finanzieller Verlust)
- Berücksichtigung einer Kostenmatrix für realitätsnahe Bewertung

---

Fazit:

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.