Dokumentation – Meilenstein 2: Datenbereitstellung

TEIL RAPHAEL

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. 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. 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. 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.