Dokumentation – Meilenstein 2: Datenaufbereitung und Datenanalyse

Vorabanalyse

Matthias

Verteilung klassifizierter Daten gegenüber nicht-klassifizierten Daten, Verteilung Trainingsdaten gegenüber Testdaten

Datentransformation

David…

-Imputationsstrategie, Feature Selektion, Transformationsschritte

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.

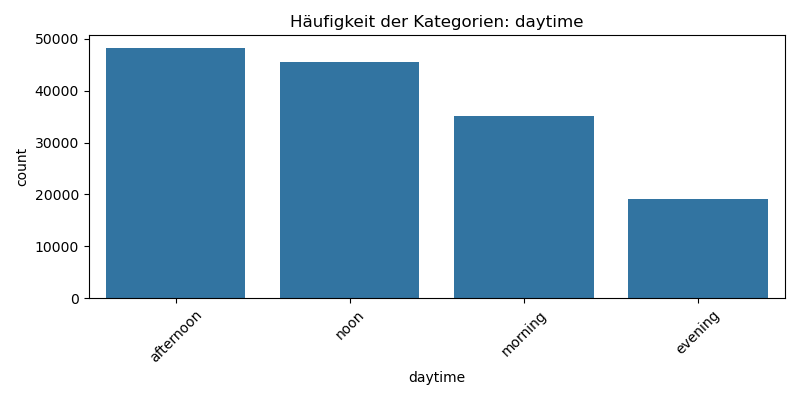
Die Vermutung liegt nahe, dass die beobachteten Extremwerte nicht als Störgröße, sondern als relevante Erklärkraft interpretiert werden können. Eine detaillierte Analyse findet sich in den folgenden Abschnitten.

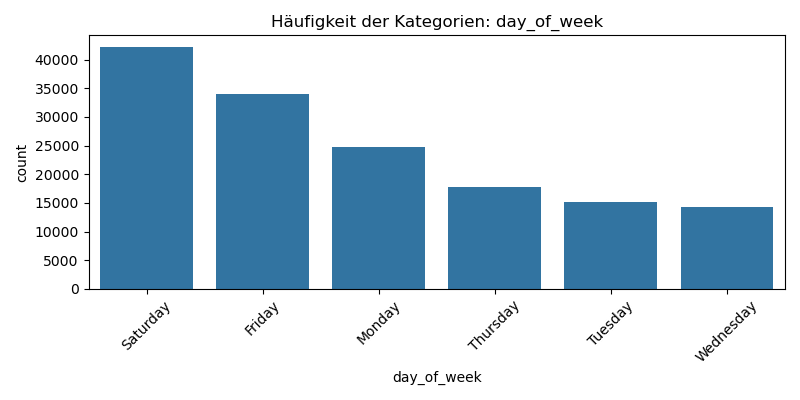
1. **Analyse kategorialer Attribute**

**Verteilung der Kategorien**

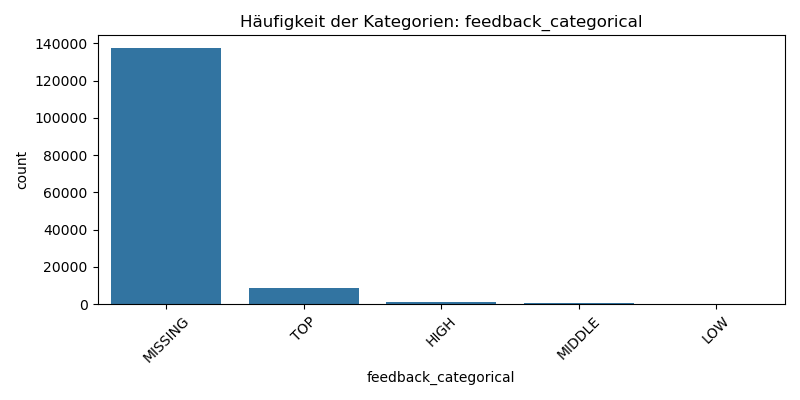
Die Häufigkeitsverteilungen aller kategorialen Variablen wurden mittels Balkendiagrammen visualisiert. Dabei zeigten sich teils starke Unausgeglichenheiten zwischen den Klassen.

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.

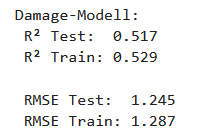
**Multivariate Regressionsanalyse**

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.

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?