Erkennung fehlerhafter Bezahlvorgänge an Selbstbedienungskassen im Einzelhandel

Datenbereitstellung

Dominik Lewin, Mario Teßmann, Johannes Winkler

7. Mai 2024

1 Ursprungsdatenquelle

1.1 Beschaffung und Verwaltungsaufwand

Die Daten sind bereits vorhanden und werden von der Auftraggeberin in einer CSV-Datei zur Verfügung gestellt. Es müssen keine zusätzlichen externen Daten erhoben oder angekauft werden. Der Beschaffungsaufwand ist demnach sehr gering.

Die Datenspeicherung erfolgt zunächst in der übergebenen CSV-Datei, da die Größe des Datensatzes gering ist. Da der Datensatz bereits anonymisiert und nicht auf Kunden zurückzuführen ist, sind bei der Datenspeicherung in Bezug auf die Vertraulichkeit und die DSGVO keine Besonderheiten zu beachten.

1.2 Datenqualität

Die Datenqualität soll im Folgenden anhand ausgewählter Kriterien bewertet werden.

1.2.1 Datenvalidität

Die Attribute beinhalten genau die Informationen, die ihre Namen versprechen. Lediglich für das Attribut *customer_feedback* findet man eine Inkonsistenz: Kunden können Sterne zwischen 1 und 5 vergeben, wobei auch halbe Sterne möglich sind. Damit sollte das Feedback ganzzahlig zwischen 1 und 10 liegen. Man findet jedoch auch Werte mit Nachkommastellen.

Es folgt eine Erklärung der einzelnen Attribute:

- GUID: Einmalige Identifikationsnummer
- grand_total: Gesamtbetrag des Einkaufs in Euro
- n_items: Anzahl gescannter Artikel
- total_checkout_time: Dauer des Scanvorgangs in Sekunden
- line_voids: Anzahl stornierter Artikel
- most_freq_product: Das am häufigsten gekaufte Produkt des Einkaufs
- products: Gescannte Produkte, bestehend aus acht verschiedenen Produktkategorien (*alcohol, fruit and vegetables, snack, dry, convenience, bakery, household, energy_drink*)
- timestamp: Genaue Zeitangabe des Einkaufs
- payment_medium: Art der Bezahlung (Karte oder Bargeld)
- label: Gibt an, ob es sich um einen Betrug (fraud) handelt oder nicht (normal)
- customer_feedback: Feedback der Kunden
- cash_desk_id: Nummer der verwendeten SB-Kasse

1.2.2 Reliabilität

Die Daten sind verlässlich, aber nur bedingt reproduzierbar. Eine erneute Erhebung der Bezahlvorgänge an SB-Kassen würde in denselben Filialen voraussichtlich sehr ähnliche Ergebnisse liefern. Eine exakte Reproduzierbarkeit ist allerdings nicht möglich, da die Einkäufe nie exakt übereinstimmen. Zudem wurden die Daten in 2017 und 2018 erhoben. Seitdem kann sich das Kundenverhalten geändert haben.

1.2.3 Repräsentativität

Die vorhandenen Daten sind über einen Zeitraum von zwei Jahren erhoben worden. Die Daten wurden in drei unterschiedlichen Filialen in Deutschland gesammelt, wobei alle Einkäufe an SB-Kassen in den Datensatz eingeflossen sind. Der Aufbau des SB-Kassen-Bereiches und das Sortiment des Einzelhandels sind in allen drei Filialen gleich. Aufgrund soziodemografischer Abweichungen in anderen Regionen kann eine Repräsentativität der vorhandene Daten allerdings nicht für alle Regionen garantiert werden. Einkäufe in anderen Filialen der Region werden sich aber vermutlich ähneln. Für die Umgebung wird die Stichprobe der drei Filialen daher als repräsentativ eingestuft, nicht hingegen für ganz Deutschland.

1.2.4 Aktualität

Die Daten wurden im Zeitraum vom 02.01.2017 bis zum 31.12.2018 erhoben. Die Daten sind somit ca. 6-7 Jahre alt. Aufgrund der Corona-Pandemie in den Jahren 2020 bis 2022 kam es durch Engpässe zu deutlichen Preissteigerungen. Weiterhin sind die Lebensmittelpreise aufgrund der aktuellen Inflation zum Teil signifikant gestiegen. Dies kann zu einem veränderten Kundenverhalten und ggf. zu einer höheren Betrugsrate geführt haben. Die Daten eignen sich demnach nur bedingt, um ein Modell zu trainieren, das auf aktuelle Einkäufe in 2024 angewendet werden soll. Da allerdings auch ein Testdatensatz aus dem Jahr 2019 zur Verfügung steht, kann die Qualität des Modells hieran bestimmt werden. Anschließend sollte das Modell mit aktuellen Daten trainiert werden, um es auch in der heutigen Zeit anwenden zu können.

1.2.5 Herkunft der Daten und Zuverlässigkeit

Die Herkunft der Daten ist bekannt. Diese wurden von der Auftraggeberin in drei der eigenen Filialen erhoben. Es handelte sich um echte Kunden und nicht lediglich um Testkunden. Die Quelle der Daten ist demnach als seriös einzustufen. Datenbesitzer ist und bleibt die Auftraggeberin. Die Daten wurden lediglich für das Projekt zur Verfügung gestellt.

1.2.6 Fehlerfreiheit

Im Datensatz sind keine Duplikate enthalten. Insbesondere beinhalten die Attribute *GUID* und *timestamp* keine Duplikate, was darauf schließen lässt, dass es sich bei jeder Zeile um einen eigenständigen Einkauf handelt. Zudem sind keine negativen Werte enthalten. Einige Attribute enthalten Ausreißer (z.B. besonders hohe Werte, die extrem vom Standard abweichen), die aber als plausibel eingestuft werden.

1.2.7 Konsistenz und Plausibilität

Es wurden einige Konsistenz- und Plausibilitätsprüfungen durchgeführt:

1. Es wurde geprüft, ob das am häufigsten gekaufte Produkt in der Spalte *most_freq_product* auch tatsächlich mit den gekauften Produkten aus der Spalte *products* übereinstimmt. Dies ist in allen Fällen gegeben. Für Einkäufe, bei denen zwei Produkte gleich häufig gekauft wurden, wurde konsistent das alphabetisch höhere Produkt ausgewählt.

2. Außerdem wurde überprüft, ob die Zeit des Einkaufs plausibel ist. Durch die Bildung des Quotienten $\frac{total\_checkout\_time}{n\_items}$ sollte überprüft werden, ob es Auffälligkeiten im Datensatz gibt. Da unerfahrene Kunden teilweise länger brauchen, scheinen die Zeiten plausibel.

3. Für stornierte Artikel wurde überprüft, ob die Anzahl an Stornierungen auch mit der Gesamtzahl gekaufter Produkte korreliert. Dies ist tatsächlich der Fall und demnach ebenfalls plausibel und konsistent.

4. Die Spalte *grand_ total* enthält größtenteils Euro-Beträge mit zwei Nachkommastellen. Allerdings gibt es auch eine geringe Anzahl an Euro-Beträgen, die drei Nachkommastellen enthalten.

5. Die Spalte *payment_ medium* enthält die Ausprägungen *credit, card und cash*. Da laut Auftraggeberin nur zwei Möglichkeiten zur Auswahl stehen, ist es hier zu inkonsistenten Einträgen gekommen.

1.2.8 Vollständigkeit

Der Datensatz enthält 104.646 Einträge. Für das Attribut *customer_feedback* sind die Daten unvollständig. Es gibt lediglich 8.348 vorhandene Einträge, demnach fehlen 96.298 Werte. Dies kann darauf zurückzuführen sein, dass nicht jeder Kunde nach dem Bezahlvorgang ein Feedback abgeben möchte und es nicht verpflichtend ist. Ansonsten sind alle Daten vollständig.

1.2.9 Detailliertheit

Der Datensatz besteht lediglich aus 12 Attributen, wovon aber auch nicht alle Attribute für die Problemstellung relevant sind. Auch die einzelnen Attribute sind teilweise nicht detailliert. Das gesamte Sortiment des Einzelhandels teilt sich in nur acht Kategorien, wobei hier keine Einzelpreise angegeben sind. Ein größerer Detailliertheitsgrad des Datensatzes würde die Vorhersagen vermutlich verbessern.

1.2.10 Umfang

Der Datensatz enthält 104.646 Einträge. Dies ist bereits eine gute Grundlage, um Muster finden zu können. Allerdings ist hier anzumerken, dass der Datensatz unausgewogen ist. Die Klasse der normalen Einkäufe ist im Gegensatz zu der Klasse mit den fehlerhaften Einkäufen ca. 20-mal größer. Bei der Analyse muss auf diese Besonderheit Rücksicht genommen werden.

2 Datenaufbereitung

Ziel der Datenaufbereitung ist es, die Daten in ein für die geplanten Analyseverfahren geeignetes Format zu bringen. Außerdem soll versucht werden, die Datenqualität zu erhöhen. In den nachfolgenden Unterabschnitten werden die Aufbereitungsschritte dokumentiert. Außerdem wird ein *Jupyter Notebook* zur Verfügung gestellt, das die genauen Aufbereitungsschritte enthält, um diese auch auf neuen Datensätzen ausführen zu können.

2.1 Dimensionsreduzierung

Irrelevante Merkmale werden aus dem Datensatz entfernt. Als irrelevant für die Problemstellung wurden die Attribute *GUID* und *customer_feedback* eingestuft. Die *GUID* enthält eine einmalige Identifikationsnummer. Diese hat keine Aussagekraft in Bezug auf einen fehlerhaften Bezahlvorgang. Auch das Feedback der Kunden hat keinen Einfluss auf die Problemstellung, insbesondere weil hier nur wenige Daten vorhanden sind. Nur in ca. 8% der Daten ist ein Kundenfeedback angegeben, was nicht repräsentativ ist. Die Einträge reichen so nicht aus und eine sinnvolle Ersetzungsstrategie für fehlende Werte wurde nicht gefunden bzw. würde die Daten verfälschen.

2.2 Merkmalserzeugung

Aus den vorhandenen Daten können weitere nützliche Attribute abgeleitet werden. Nachfolgend sollen die Attribute näher erläutert werden, aus denen sich neue Attribute extrahieren lassen.

2.2.1 *timestamp*

Aus dem Attribut *timestamp* werden die Attribute *weekday* und *hour* erzeugt, da hier ein Zusammenhang zum Label gegeben ist. Beim Wochentag häufen sich die Fälle fehlerhafter Bezahlvorgänge durchschnittlich an Freitagen und Samstagen, vgl. Anhang A, Abbildung 1. Bei der Uhrzeit (volle Stunde) fällt auf, dass in der Zeit zwischen 16 und 19 Uhr die meisten fehlerhaften Bezahlvorgänge zu finden sind, vgl. Anhang A, Abbildung 2. Allerdings ist hier anzumerken, dass sich die Einkäufe an den genannten Wochentagen und Zeiten generell erhöhen. Relativ gesehen, sind hier kaum Unterschiede zu erkennen.

Der Timestamp selbst wird anschließend aus dem Datensatz entfernt, da dieser viele nicht benötigte Informationen enthält. Weitere Attribute, die sich aus dem Timestamp ableiten lassen (Jahr, Monat, Tag, Minute), hängen nicht mit der Klassifikation zusammen.

Anhang A, Abbildung 3 kann entnommen werden, dass es zwischen den einzelnen Monaten keine nennenswerten Unterschiede zwischen normalen und fehlerhaften Bezahlvorgängen gibt. Da auch in absoluten Zahlen keine Besonderheiten auffallen, wird dieses Attribut nicht verwendet.

Für den Tag des Monats wurde zunächst angenommen, dass zum Ende eines Monats ein Anstieg fehlerhafter Bezahlvorgänge zu erwarten ist. Diese Vermutung spiegelte sich in den Daten nicht wider. Stattdessen ist im Monatsverlauf kein Muster zu erkennen, vgl. Anhang A, Abbildung 4. Aus diesem Grund wird auch dieses Attribut nicht verwendet.

Die Minuten werden nicht berücksichtigt, da die Stunde bereits enhalten ist und eine gute Eingrenzung der Tageszeit liefert. Die Minuten würden lediglich zu einer Überanpassung des Modells führen. Gleiches gilt auch für das Jahr. Da die Vorhersagen auf einem aktuelleren Zeitraum laufen sollen, wird dieses Attribut nicht gebraucht.

2.2.2 *products*

Das Attribut *products* enthält eine Zeichenkette, die die einzelnen gescannten Produkte (bzw. deren Produktkategorie) der Reihenfolge nach enthält. Da die Reihenfolge irrelevant und lediglich die Anzahl der im Einkauf enthaltenen Produkte von Bedeutung ist, kann für jede Produktkategorie eine eigene Spalte im Datensatz aufgenommen werden, die die Anzahl enthält. Für die acht Kategorien werden somit neue Attribute erzeugt. Sollte eine Produktkategorie nicht in einem Einkauf vorkommen, wird der Eintrag auf 0 gesetzt. Das Attribut *products* wird anschließend nicht mehr gebraucht und aus dem Datensatz entfernt.

2.3 Datenbereinigung

Während der Datenbereinigung wurden identifizierte Fehler bereinigt und Datentypen zur besseren Verarbeitung verändert. Dies betrifft die folgenden Attribute:

- *checkout_ time*: Der Detailiertheitsgrad ist zu hoch, da die Milisekunden für die Problemstellung irrelevant sind. Die Werte wurden auf ganze Zahlen gerundet.
- *grand_ total*: Alle Einträge, die drei Nachkommastellen hatten, wurden auf zwei Nachkommastellen gekürzt. Dies entspricht den restlichen Daten und den tatsächlichen Einkaufspreisen im Einzelhandel.
- *payment_ medium*: Die Zahlungsart Karte wurde bis zum 12.08.2017 für fehlerhafte Daten zunächst als *card* gespeichert. Für den restlichen Zeitraum bis zum 31.12.2018 wurde dieselbe Zahlungsart als *credit* gespeichert. Um diesen inkonsistenten Eintrag zu bereinigen, wurden alle Kartenzahlungen einheitlich mit *card* überschrieben.

2.4 Datentransformation

Um die Daten schließlich noch für das Training verschiedener Modelle vorzubereiten, müssen einige Attribute transformiert werden. Bei *most_freq_ product* und *payment_ medium* handelt es sich um kategorische Attribute. Diese könnten nun entweder durch Ganzzahlen oder One-Hot-Codierung codiert werden. Ersteres ist für Machine-Learning-Algorithmen nur sinnvoll, wenn die Attribute ordinal sind, was nicht der Fall ist. Aus diesem Grund werden die genannten Attribute one-hot-codiert.

Das Attribut *label* wird als binäres Attribut durch die Zahlen 0 und 1 dargestellt, wobei 0 für *normal* und 1 für *fraud* steht.

Das Attribut *weekday* ist durch die Auswertung des timestamps bereits als Ganzzahl gespeichert worden. Die Wochentage Montag bis Samstag sind durch die Zahlen 0 bis 5 dargestellt. Da dieses Attribut ordinal ist, also eine feste Reihenfolge besteht, ist eine Transformation in one-hot zunächst nicht notwendig, kann aber im Laufe der Analyse angepasst werden, um ggf. bessere Ergebnisse erzielen zu können.

2.5 Protokollierung der Datenaufbereitung

Zur besseren Nachvollziehbarkeit, zur Reproduzierbarkeit der Ergebnisse und für das Trainieren der Modelle mit Hilfe neuer Daten wird ein *Jupyter Notebook* zur Verfügung gestellt. Dieses liest

den originalen Datensatz ein, verarbeitet die Daten entsprechend der vorliegenden Dokumentation und speichert schließlich den aufbereiteten Datensatz.

3 Datenmanagement

Da, wie erwähnt, die Auftraggeberin Besitzerin der Daten ist und bleibt, obliegt ihr das Management der bereits vorhandenen Daten. Insbesondere sind die Analyseverfahren jederzeit reproduzierbar, da sie im Besitz der Trainingsdaten ist. Im weiteren Verlauf des Projekts wird beschrieben, wie mit neu zu erhebenden Daten umgegangen werden kann.

Die Ursprungsdaten bleiben unberührt. Da die Daten im Laufe des Projekts nicht ansteigen werden und die Speichergröße des Datensatzes gering ist, ist eine zusätzliche Infrastruktur zur Speicherung zum aktuellen Zeitpunkt nicht notwendig. Die Datensicherung der aufbereiteten Daten erfolgt zunächst ebenfalls aufgrund der geringen Größe ohne gesonderte Infrastruktur in einer CSV-Datei.

4 Explorative Datenalayse

Die Ergebnisse der explorativen Datenanalyse werden gesondert in einer Präsentation am 07.05.2024 vorgestellt, auf die an dieser Stelle verwiesen sei. Ergänzend zu den hier dargestellten Inhalten (Datenvalidierung, inhaltliches Verständnis und Identifikation zentraler Attribute), die aufgrund der engen Verbundenheit der einzelnen Phasen ebenfalls Teil der explorativen Datenanalyse sind, werden wir in der Präsentation auf die Datenvisualisierung, statistische Maße und die Ausreißeridentifikation eingehen. Die Präsentationsfolien werden im Nachgang zur Verfügung gestellt und vervollständigen dieses Dokument.

A Anhang

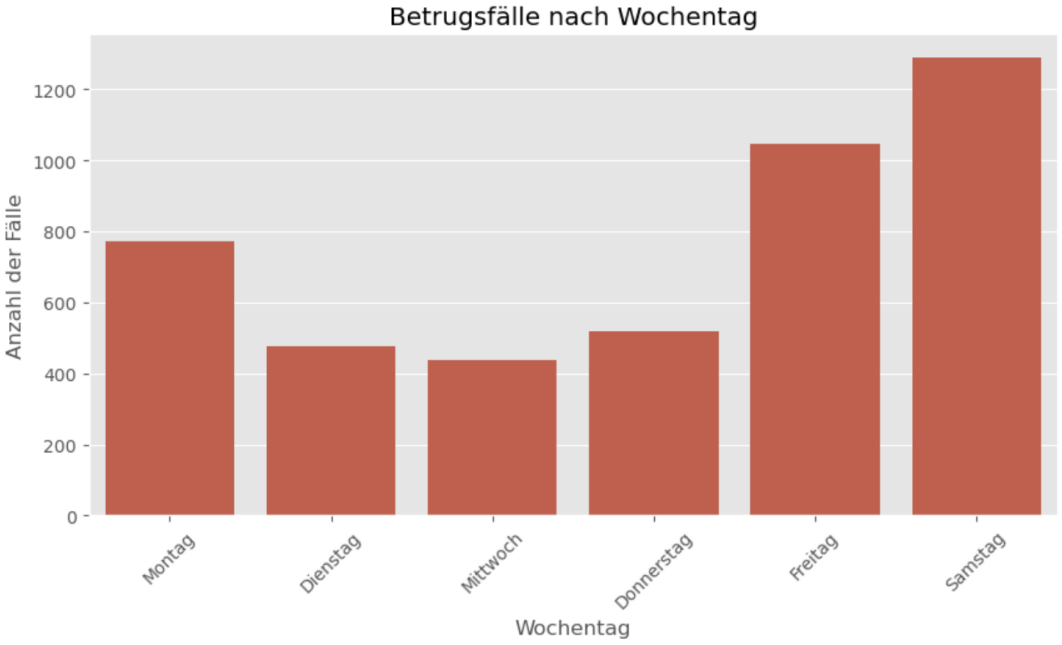

Abbildung 1: Anzahl fehlerhafter Bezahlvorgänge pro Wochentag

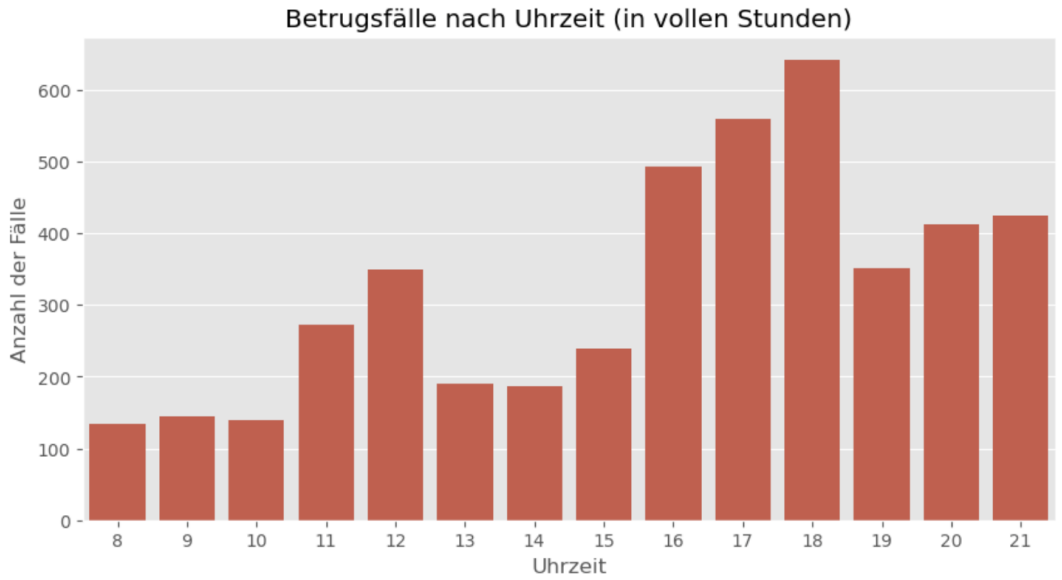

Abbildung 2: Anzahl fehlerhafter Bezahlvorgänge nach Uhrzeit (in vollen Stunden)

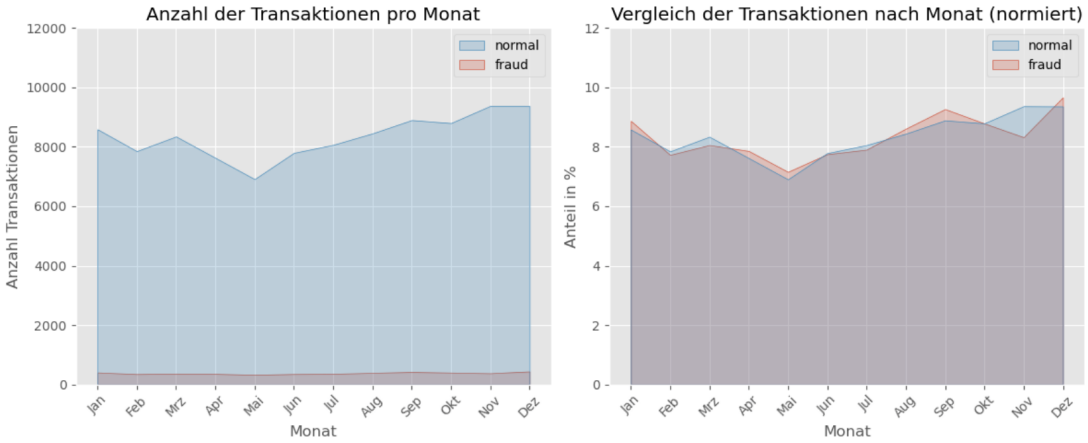

Abbildung 3: a) Anzahl korrekter und fehlerhafter Transaktionen nach Monat. b) Vergleich korrekter und fehlerhafter Transaktionen nach Monat, jeweils normiert auf 100 %.

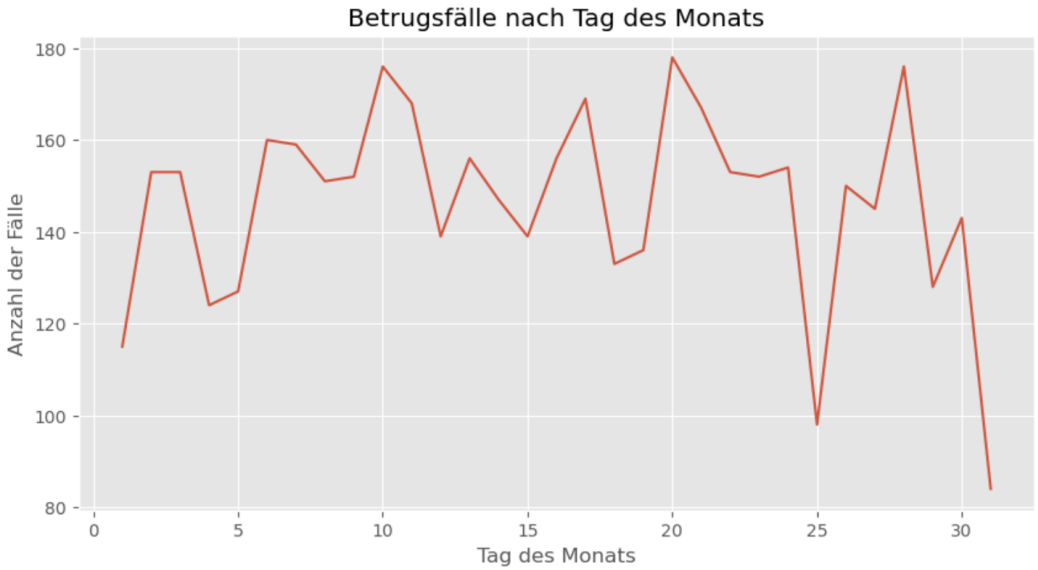

Abbildung 4: Anzahl fehlerhafter Bezahlvorgänge im Monatsverlauf nach Tagen