UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aplicação de aprendizado de máquina na identificação de erros atípicos nas transações de envio do Pix

Rosely Nakaza

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Aplicação de aprendizado de máquina na identificação de erros atípicos nas transações de envio do Pix

---

Rosely Nakaza

USP - São Carlos
2025

Rosely Nakaza

Aplicação de aprendizado de máquina na identificação de erros atípicos nas transações de envio do Pix

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial.

Orientador: Me. Willian Dener de Oliveira

USP - São Carlos
2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

NAKAZA, ROSELY
NN163a Aplicação de aprendizado de máquina na identificação de erros atípicos nas transações de envio do Pix / ROSELY NAKAZA; orientador Willian Dener de Oliveira. -- São Carlos, 2025.
51 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Pix. 2. Detecção de Anomalias. 3. Isolation Forest. 4. Decomposição STL. 5. Séries Temporais. I. de Oliveira, Willian Dener, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Dedico este trabalho à minha mãe, um ser humano inspirador de luz e amor, a quem devo toda a base e força para buscar conhecimento.

AGRADECIMENTOS

Agradeço, em especial, ao meu esposo Daniel, que me apoiou incansavelmente durante todo o MBA.

Ao meu orientador, Me. Willian Dener de Oliveira, pelos seus direcionamentos, paciência, disponibilidade e bom humor em nossas reuniões.

Agradeço às ferramentas de Inteligência Artificial, Manus e Gemini, que auxiliaram na organização da estrutura do trabalho, correção de códigos e gráficos em Python e revisão de textos.

E aos meus colegas Rodrigo Ortega e Rodrigo Teles, que me incentivaram e ajudaram a desenvolver o modelo de detecção de anomalias contido neste trabalho.

RESUMO

NAKAZA, R. **Aplicação de aprendizado de máquina na identificação de erros atípicos nas transações de envio do Pix**. 2025. 51 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este trabalho aborda a necessidade de detecção proativa de anomalias em sistemas de pagamentos instantâneos, com foco nos erros de transações do Pix, o principal meio de pagamento do país. O objetivo principal foi desenvolver e validar uma metodologia para identificar padrões incomuns que possam indicar falhas operacionais ou desvios críticos no comportamento transacional. Para tanto, empregou-se uma abordagem que combina a decomposição STL (*Seasonal-Trend decomposition using Loess*) para séries temporais, permitindo a separação dos componentes de tendência, sazonalidade e resíduos dos dados de erros, com o algoritmo *Isolation Forest*, aplicado aos resíduos para a detecção eficiente de anomalias.A metodologia foi aplicada a um conjunto de dados reais de erros de transações Pix de uma instituição financeira, abrangendo o período de outubro a dezembro de 2024. Os resultados demonstraram a eficácia do modelo na identificação de eventos anômalos, que foram posteriormente contextualizados e validados por especialistas da área de TI e gestão. A análise detalhada de dias com maior ocorrência de anomalias, como 29 de novembro (*Black Friday*) e 6 de dezembro (período de alto volume devido ao pagamento de salários/gratificações e proximidade das festas de fim de ano), revelou a capacidade do modelo de discernir entre desvios causados por picos de demanda e regras de negócio (limites noturnos do Pix) e falhas no próprio sistema de telemetria, atuando como um mecanismo de *data observability*. A validação humana mostrou-se indispensável para a correta interpretação e diferenciação entre anomalias críticas e variações operacionais esperadas, minimizando falsos positivos. Conclui-se que a metodologia proposta oferece uma ferramenta para o monitoramento contínuo e a gestão proativa da qualidade e disponibilidade do serviço Pix, permitindo a identificação precoce de potenciais problemas e o aprimoramento da infraestrutura. As contribuições deste trabalho incluem uma metodologia integrada e validada, uma análise contextual aprofundada e a demonstração da capacidade de detecção de falhas de telemetria, pavimentando o caminho para futuros estudos em automação da validação e adaptação a ambientes de tempo real.

Palavras-chave: Pix. Detecção de Anomalias. *Isolation Forest*. Decomposição STL. Séries Temporais. *Data Observa*bility.

ABSTRACT

NAKAZA, R. **Machine learning application to identify atypical errors in Pix Sending Transactions**. 2025. 51 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

This work addresses the need for proactive anomaly detection in instant payment systems, with a focus on Pix transaction errors, the country's primary payment method. The primary objective was to develop and validate a methodology for identifying unusual patterns that may indicate operational failures or critical deviations in transaction behavior. To this end, an approach was employed that combines STL (Seasonal-Trend Decomposition using Loess) decomposition for time series, allowing the separation of trend, seasonality, and residual components of error data, with the Isolation Forest algorithm applied to the residuals for efficient anomaly detection. The methodology was applied to a real dataset of Pix transaction errors from a financial institution, covering the period from October to December 2024. The results demonstrated the model's effectiveness in identifying anomalous events, which were subsequently contextualized and validated by IT and management experts. A detailed analysis of days with the highest occurrence of anomalies, such as November 29th (Black Friday) and December 6th (a period of high volume due to salary/bonus payments and the proximity of the holiday season), revealed the model's ability to distinguish between deviations caused by peak demand and business rules (nighttime Pix limits) and failures in the telemetry system itself, acting as a data observability mechanism. Human validation proved essential for the correct interpretation and differentiation between critical anomalies and expected operational variations, minimizing false positives. We conclude that the proposed methodology offers a tool for continuous monitoring and proactive management of Pix service quality and availability, enabling the early identification of potential problems and infrastructure improvements. The contributions of this work include an integrated and validated methodology, an in-depth contextual analysis, and the demonstration of telemetry failure detection capabilities, paving the way for future studies on validation automation and adaptation to real-time environments.

Keywords: Pix. Anomaly Detection. Isolation Forest. STL Decomposition. Time Series. Data Observability.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Criado pelo Banco Central do Brasil (BACEN) em 2020, o Pix tornou-se rapidamente o principal meio de pagamento dos brasileiros. O sistema, que permite transferências instantâneas entre contas 24 horas por dia, é a modalidade mais utilizada no país, com adesão de 76,4% da população, segundo pesquisa da própria instituição. Em 2024, foram realizadas 63,4 bilhões de transações, com volume total de R$ 26,4 trilhões (BACEN, 2024a, 2024c).

Para garantir a operacionalidade do grande volume, o Pix requer um alto nível de estabilidade, princípio fundamental da infraestrutura do mercado financeiro (CPMI; IOSCO, 2012). Apesar de ser projetado para ter alta disponibilidade, não está imune a falhas. Erros de transação, mesmo que raros, podem ocorrer, sejam de origem técnica ou operacional, um conceito alinhado aos princípios da Engenharia de Confiabilidade de Sistemas (BEYER et al., 2016). A própria regulamentação do Pix pelo Banco Central (2020d) prevê diversos motivos de falha, cujas mensagens de erros estão detalhadas nos Requisitos Mínimos para Experiência do usuário (BACEN, 2025a), reforçando a necessidade de um monitoramento contínuo e inteligente para detectar padrões anômalos e garantir a integridade do ecossistema.

Quando esses erros ocorrem em padrões incomuns, como um aumento súbito em um tipo específico de falha, eles podem ser o sintoma de um problema maior, com potencial para impactar a experiência dos usuários, gerar custos operacionais e penalidades e multas previstas para as instituições envolvidas (BACEN, 2021b).

Uma das formas para identificar o comportamento atípico em grande volume de dados é a utilização de técnicas de aprendizado de máquina. Uma abordagem aplicável a dados sequenciais, como os erros em transação do Pix, consiste em utilizar no pré-processamento a decomposição de séries temporais. Esta técnica permite remover padrões previsíveis de tendência e sazonalidade (Cleveland et al., 1990) para que, em seguida, sejam aplicados sobre esses resíduos algoritmos de detecção de anomalias, como o *Isolation Forest* – Floresta de Isolamento (LIU; TING; ZHOU, 2008).

1.1 Justificativa e motivação

No Brasil, onde o Pix é o principal meio de pagamento, qualquer falha, mesmo que momentânea, pode gerar frustração e causar prejuízos a pessoas físicas e jurídicas, bem como abalar a credibilidade e a satisfação nas instituições financeiras, motivando o cliente a migrar para um concorrente.

Para as instituições financeiras, a importância da detecção proativa de anomalias nos erros de transação vai além da simples monitoração técnica, sendo uma ferramenta para gestão de risco e governança. A identificação de padrões incomuns permite prevenir incidentes críticos, auxiliando na tomada de ações nos primeiros sinais de problemas sistêmicos ou possíveis ameaças à segurança, antes que elas escalem. Adicionalmente, mitiga riscos financeiros e regulatórios, reduzindo a exposição a perdas operacionais, evitando a aplicação de penalidades e multas previstas pelo Banco Central (BACEN, 2021b) e contribui para a preservação da imagem da marca, evitando a exposição negativa e o risco de imagem associado a falhas.

Dessa forma, o desenvolvimento de um modelo de detecção de anomalias é mais do que um aprimoramento técnico, é um investimento na resiliência operacional, na conformidade regulatória e na sustentabilidade do negócio no longo prazo.

1.2 Objetivos

O objetivo deste trabalho é desenvolver um modelo para detecção de anomalias em erros de transação de envio do Pix, no contexto de uma instituição financeira, combinando a decomposição de séries temporais com o algoritmo *Isolation Forest*. A partir do objetivo geral, foram definidos os seguintes objetivos específicos:

- Coletar e pré-processar um conjunto de dados de erros de transações Pix.
- Aplicar a decomposição de séries temporais para separar os componentes de sazonalidade, tendência e resíduo da série temporal.
- Treinar o modelo *Isolation Forest* utilizando o componente de resíduo da decomposição.
- Avaliar os resultados do modelo, com a análise da relevância das anomalias detectadas por especialistas no domínio.

Assim, espera-se contribuir na tomada de medidas corretivas, se for o caso, de forma proativa e tempestiva, evitando interrupções nas operações do Pix, prejuízos e impactos negativos junto aos clientes e à imagem do Banco.

1.3 Estrutura do trabalho

Este trabalho foi organizado da seguinte forma: no capítulo 2 é apresentada a fundamentação teórica com a descrição sobre o Pix, análise e decomposição de séries temporais, detecção de anomalias e o algoritmo *Isolation Forest*. O capítulo 3 descreve a metodologia utilizada, incluindo o pré-processamento, a aplicação da decomposição STL e o treinamento do algoritmo *Isolation F*orest. O capítulo 4 foca na análise dos resultados e na discussão sobre as anomalias encontradas. Por fim, no capítulo 5 são apresentadas as conclusões e sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

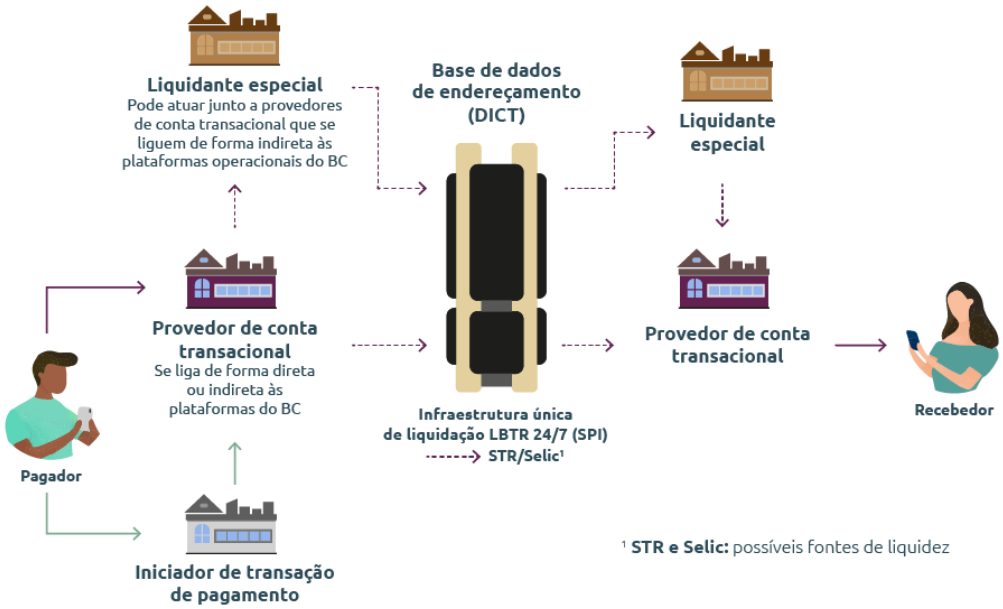
2.1 O sistema de pagamentos instantâneo Pix

O Pix é um meio de pagamento instantâneo criado pelo Banco Central do Brasil que possibilita a transferência entre contas em poucos segundos e pode ser feita a qualquer hora do dia (BACEN, 2020c). Rápido, simples, acessível e seguro, em apenas quatro anos o sistema consolidou-se como a principal ferramenta para envio e recebimento de recursos pelos brasileiros (BACEN, 2024c), sendo uma alternativa para as transferências bancárias tradicionais, pagamentos em dinheiro e com cartão.

2.2 Arquitetura e funcionamento

O Banco Central do Brasil atua como regulador, gestor e operador do Pix, sendo responsável tanto pela definição de suas regras de funcionamento quanto pelo desenvolvimento de sua infraestrutura de liquidação e plataforma tecnológica (BACEN, 2020b). Abaixo, a Figura 1 mostra o processo de funcionamento do Pix com o detalhamento dos principais componentes do seu ecossistema.

Figura 1 – Ecossistema de pagamentos instantâneos brasileiro



Fonte: BACEN (2020b).

Pagador: acessa a conta em sua instituição financeira ou instituição de pagamento participante do Pix e inicia a transação.

Provedores de contas transacionais: As instituições financeiras e de pagamentos que oferecem o serviço diretamente ao usuário final e se conectam ao Sistema de Pagamentos Instantâneos (SPI) para efetivar as operações.

Liquidante especial: instituição financeira ou de pagamento autorizada a funcionar pelo BACEN e que presta, exclusivamente, serviço de liquidação para outros participantes do Pix que não acessam o Sistema de Pagamentos Instantâneos (SPI).

Iniciador de transação de pagamento: quando a instituição que inicia o pagamento é diferente daquela que detém a conta do usuário pagador. Ocorre, por exemplo, quando o usuário faz uma compra online em uma loja e é direcionado para a tela de autenticação da transação no aplicativo do seu banco. Após a confirmação, é redirecionado de volta à loja virtual (BACEN, 2021a).

Sistema de Pagamentos Instantâneos (SPI): A infraestrutura central e única de liquidação que processa as transações do Pix (BACEN, 2025b).

Diretório de Identificadores de Contas Transacionais (DICT): Banco de dados que armazena as informações cadastrais dos usuários recebedores e contas transacionais como as "Chaves Pix" (CPF/CNPJ, número de celular, e-mail ou chave aleatória). Permite a iniciação do pagamento de forma simples e segura (BACEN, 2020a).

Para este trabalho, os erros estudados serão de uma transação de envio do Pix que ocorrem em um provedor de contas transacionais, dentro do sistema de processamento transacional.

2.3 Análise de séries temporais

Uma série temporal é uma sequência de observações ou pontos de dados coletados em tempos regulares. A análise de séries temporais tem como objetivo investigar o mecanismo gerador da série temporal, fazer previsões de valores futuros da série, descrever o comportamento da série e procurar periodicidades relevantes nos dados (MORETTIN; TOLOI, 2018). Neste trabalho, a base de dados utilizada contém a contagem de erros por transação a cada 15 minutos, constituindo uma série temporal. A análise focará no estudo do comportamento da série como tendência e variações sazonais e resíduos, extraídos com a decomposição conforme veremos a seguir.

Uma série temporal pode ser decomposta em até quatro componentes principais que, quando combinados, reconstituem a série original:

1. **Tendência (T):** Refere-se à direção ou ao movimento de longo prazo da série. Uma tendência pode ser crescente (aumento gradual de erros devido ao crescimento do uso da transação), decrescente ou estável.
2. **Sazonalidade (S):** Flutuações periódicas e previsíveis que ocorrem em intervalos fixos de tempo. Em dados de transações financeiras, a sazonalidade costuma ser um componente relevante pois segue o comportamento dos usuários, geralmente manifestando-se em padrões diários, com picos em horário comercial e semanais, com diferenças entre dias úteis e fins de semana, por exemplo.
3. **Ciclo (C):** Corresponde a flutuações de longo prazo que não possuem um período fixo, geralmente associadas a ciclos econômicos.
4. **Resíduo (R):** Também conhecido como erro ou ruído, o resíduo é a variação irregular que permanece após a remoção dos componentes de tendência e sazonalidade. Representa os eventos não sistemáticos e imprevisíveis da série.

A relação entre esses componentes pode ser modelada de forma aditiva (Y = T + S + R), quando a magnitude da sazonalidade é constante ao longo do tempo, ou multiplicativa (Y = T x S x R), quando a variação sazonal aumenta ou diminui proporcionalmente ao nível da série (HYNDMAN; ATHANASOPOULOS, 2018). No caso dos erros em transações de envio do Pix a modelagem é aditiva, uma vez que as flutuações sazonais na quantidade de erros permanecem constantes ao longo do tempo, independentemente do nível da tendência, conforme detalhado no Capítulo 3.3.4 – Análise Exploratória.

2.4 Decomposição de séries temporais

Para separar os componentes da série temporal de erros, este trabalho utiliza o método de decomposição sazonal e de tendência usando *Loess* (STL), proposto por Cleveland et al.(1990). O método utiliza a técnica de regressão local ponderada, conhecida como *Loess* (*Locally Estimated Scatterplot Smoothing)*, para suavizar a série e extrair seus componentes de forma iterativa. O processo consiste em um conjunto de laços internos e externos. O laço interno alterna entre a estimação da tendência e dos ciclos sazonais, enquanto que o laço externo calcula pesos de robustez para reduzir a influência de observações atípicas no processo de estimação.

As principais vantagens do STL são:

- **Robustez a outliers:** O uso do *Loess* e dos pesos de robustez torna o STL menos sensível a valores extremos, permitindo que o método isole a tendência e a sazonalidade sem ser distorcido pelas próprias anomalias que se deseja detectar.

- **Flexibilidade:** O STL permite que o componente sazonal mude ao longo do tempo.
- **Versatilidade:** Pode ser aplicado em séries com qualquer tipo de padrão sazonal (diário, semanal, mensal) e não se limita a dados sem valores ausentes.

Ao final do processo, o STL fornece três séries distintas: uma com a tendência, uma com a sazonalidade e uma com o resíduo. O modelo para detecção de anomalias será focado na análise de resíduos, para que o algoritmo se concentre nas variações que não são explicadas pelos padrões normais de erros nas transações.

2.5 Detecção de anomalias

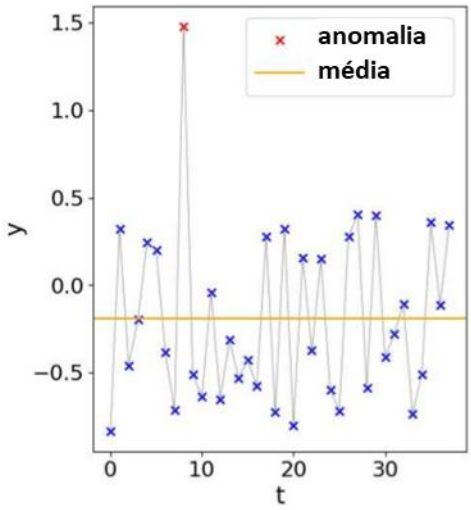
Segundo Chandola, Banerjee e Kumar (2009), a detecção de anomalias refere-se a encontrar padrões em dados que não estão em conformidade com o comportamento esperado. Esses padrões são frequentemente chamados de anomalias, *outliers*, observações discordantes, exceções, aberrações, surpresas, peculiaridades ou contaminantes em diferentes domínios de aplicação.

Ahmed, Mahmood e Islam (2016) destacam que essas anomalias são importantes porque indicam eventos significativos, mas raros, que podem levar à tomada de decisões em diversos setores. No contexto de análise de erros nas transações do Pix, podem identificar desde uma tentativa de fraude até uma falha técnica, contribuindo para a tomada de ações corretivas para manutenção da segurança e disponibilidade da transação.

De acordo com Chandola, Banerjee e Kumar (2009), as anomalias podem se enquadrar em três tipos:

1. **Anomalia pontual**: quando uma instância de dados individual é considerada anômala em relação ao restante dos dados, como pode ser observado na Figura 2. Outliers pontuais são pontos de dados que se desviam individualmente dos pontos de dados vizinhos. No escopo deste trabalho, um pico súbito e isolado na quantidade de erros em um curto intervalo de tempo seria um exemplo de anomalia pontual.

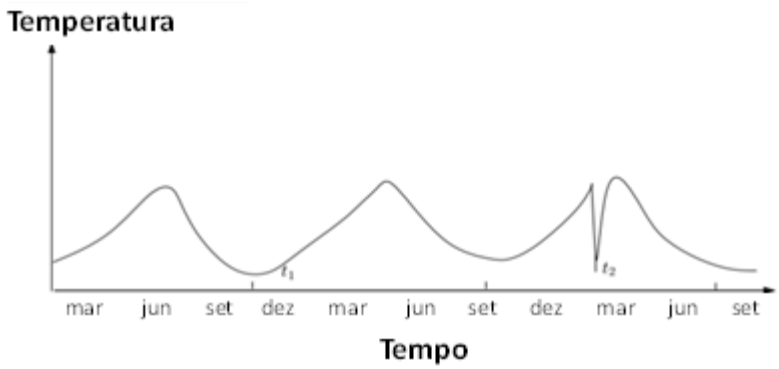
Figura 2 – Outliers pontuais



Fonte: Schindler, Schlicht, Thoben (2023)

2. **Anomalias contextuais ou condicionais**: quando uma instância de dados é anômala em um contexto específico, mas não de outra forma. O contexto depende da estrutura no conjunto de dados e deve ser especificado como parte da formulação do problema, conforme a Figura 3.

Figura 3 – Anomalia contextual t2 em uma série temporal de temperatura

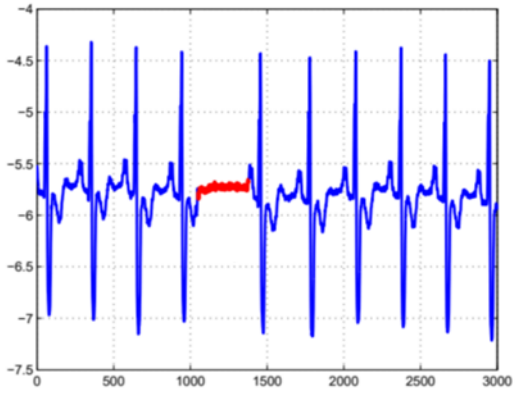


Fonte: Chandola, Banerjee e Kumar (2009)

No gráfico, a temperatura em t1 é a mesma que t2, mas ocorre em um contexto diferente e, por isso, não é considerada uma anomalia. Para os erros nas transações Pix, por exemplo, um volume de 1.000 ocorrências pode ser normal em um horário de pico comercial, mas seria anômalo se ocorresse de madrugada, caracterizando uma anomalia contextual.

3. **Anomalias coletivas**: quando uma coleção de dados relacionados for anômala comparativamente ao conjunto de dados inteiro, embora os dados individuais possam não ser anômalos por si só, conforme ilustrado na Figura 4. A região destacada revela uma anomalia uma vez que o mesmo valor baixo ocorre por um período de tempo anormalmente longo. O valor baixo por si só não é uma anomalia.

Figura 4 – Eletrocardiograma humano.



Fonte: Chandola, Banerjee e Kumar (2009)

Um exemplo para detecção de anomalias em erros de transação seria um aumento pequeno, mas sustentado, no número de erros por período prolongado, que desvia da tendência esperada.

Existem diversas abordagens para detecção de anomalias, que podem ser agrupadas em categorias como métodos estatísticos, por exemplo, *Z-score* e Gráfico de Controle, baseados em proximidade, como *K-Nearest Neighbors* e *Local Outlier Factor* – LOF e baseados em aprendizado de máquina, como o *Isolation Forest*, que será utilizado neste estudo (AGGARWAL, 2017).

2.6 *Isolation Forest* (Floresta de Isolamento)

O *Isolation Forest* é um algoritmo de aprendizado de máquina não supervisionado proposto por Liu, Ting e Zhou (2008), especificamente projetado para detecção de anomalias. Sua eficácia baseia-se no princípio de que as anomalias são poucas e diferentes e, portanto, são mais suscetíveis ao isolamento do que os pontos normais.

Ao contrário de outros algoritmos que tentam construir um “perfil de normalidade” e depois identificar o que se desvia dele, o *Isolation Forest* busca ativamente isolar as anomalias em dois estágios: Treinamento e Avaliação, descritos a seguir.

Treinamento: Neste estágio, o algoritmo constrói um conjunto ou “floresta” de árvores de decisão binárias. Para cada árvore, uma amostra aleatória dos dados é selecionada. Isso é importante pois reduz o *swamping* (alagamento), que ocorre quando instâncias normais estão muito próximas das anomalias, dificultando a distinção, e o *masking* (mascaramento) que acontece quando há muitas anomalias, escondendo a sua própria presença. A subamostragem reduz a probabilidade desses efeitos, tornando as anomalias mais fáceis de isolar.

A partir da raiz da árvore, os dados são recursivamente particionados. A cada nó, o algoritmo seleciona aleatoriamente um atributo (*feature*) e um valor de divisão aleatório entre o mínimo e o máximo daquele atributo. Os dados são divididos em dois galhos – pontos com valor menor que o ponto de divisão vão para um lado e os com valor maior ou igual vão para o outro. Esse processo de particionamento continua até que cada ponto de dados da amostra esteja isolado em um nó da folha da árvore ou até que a árvore atinja uma altura máxima predefinida.

Avaliação: No estágio de avaliação, a detecção da anomalia deriva do comprimento do caminho (*path length)*, que é o número de arestas que um ponto percorre da raiz até um nó da folha. Pontos anômalos, por serem diferentes, tendem a ser isolados mais perto da raiz da árvore, resultando em um caminho mais curto. Pontos normais, por estarem agrupados e serem semelhantes entre si, exigem mais partições para serem isolados, resultando em caminhos mais longos.

Uma pontuação (*score*) de anomalia de um ponto é calculada com base no comprimento médio de seu caminho em toda a floresta de árvores. A pontuação é projetada para que valores próximos a 1 indiquem anomalias fortes. Valores muito menores que 0,5 indicam instâncias normais. Valores em torno de 0,5 indicam que não há anomalias claras na amostra. As instâncias são ranqueadas em ordem decrescente com base na sua pontuação de anomalia. Aquelas com pontuações mais altas são consideradas mais prováveis de serem anomalias.

O *Isolation Forest* possui vantagens significativas, que contribuem para a análise neste trabalho:

Eficiência computacional: Possui complexidade de tempo linear e baixo consumo de memória, sendo altamente escalável para grandes volumes de dados, como os gerados pelo Pix.

Robustez: Não requer a normalização dos dados e não faz suposições sobre a sua distribuição.

Eficácia comprovada: É um dos algoritmos de linha de base mais fortes para detecção de anomalias não supervisionada, superando métodos baseados em distância em diversos cenários (LIU; TING; ZHOU, 2008).

Com base na fundamentação teórica apresentada, que abrangeu desde a arquitetura do Pix até os princípios da análise de séries temporais e dos algoritmos de detecção de anomalias, foi delineada uma estratégia metodológica. A abordagem proposta consiste em um processo de duas etapas: primeiramente, a aplicação da decomposição STL para isolar os componentes não previsíveis da série de erros e, em segundo lugar, a utilização do algoritmo *Isolation Forest* sobre os resíduos para identificar as anomalias. O capítulo seguinte detalhará o passo a passo da implementação desta metodologia, desde o pré-processamento dos dados até a avaliação final do modelo.

3 METODOLOGIA

3.1 Visão geral da metodologia

Neste trabalho, seguimos um fluxo estruturado partindo da obtenção e preparação dos dados até a avaliação de um modelo de aprendizado de máquina não supervisionado. A abordagem foi baseada no trabalho de Glazkov (2023) e consiste em um processo com duas etapas principais: a decomposição da série temporal de erros utilizando o método STL para isolar o componente residual e, em seguida, a aplicação do algoritmo *Isolation Forest* sobre esses resíduos para identificar observações anômalas. Todas as etapas foram desenvolvidas utilizando a linguagem de programação *Python* e bibliotecas como Pandas, *Statmodels* e *Scikit-learn*.

3.2 Descrição do conjunto de dados (*dataset*)

Os dados utilizados no desenvolvimento do trabalho foram extraídos do sistema de transação de uma instituição financeira, onde são processadas as operações de movimentação das contas dos clientes. Um módulo de captura de erros na transação é acionado a cada 15 minutos, abastecendo um banco de dados que sumariza os erros por transação, canal, código do erro e posição do erro no programa. O período de extração considerado foi de 90 dias, de 2 de outubro a 30 de dezembro de 2024. A estrutura original do conjunto de dados possui as seguintes colunas:

Quadro 1 – Estrutura original do conjunto de dados com a quantidade de erros na transação de envio do Pix

Nome da coluna	Descrição
CD_TRAN	Código da transação
CD_TIP_CNL	Código do tipo de canal em que ocorreu o erro
CD_ERRO	Código do erro
NR_LINHA_ERRO_TRAN	Nº da posição do erro no programa
DT_MVT_PREV	Data da coleta dos dados

HR_MVT_PREV	Hora da coleta dos dados
QT_ERRO	Quantidade de erros
TX_MSG_ERRO	Mensagem de erro
NR_PREV	Número da prévia
DT_CRIC_ARQ	Data de criação do arquivo

Fonte: Elaborado pela autora (2025).

3.3 Pré-processamento

Neste tópico descreveremos as etapas de pré-processamento executadas na base de dados original.

3.3.1 Anonimização e padronização

A primeira etapa do pré-processamento foi o tratamento do campo TX_MSG_ERRO, que traz a descrição textual do motivo da falha na transação, por exemplo, "Saldo Insuficiente". Observamos que, para certos tipos de erros, o texto padrão vinha concatenado com dados variáveis da transação, alguns dos quais pessoais, como a chave de identificação do cliente, ou bancários, como o identificador (ID) da transação de envio do Pix. Apesar de não serem considerados dados sensíveis pela Lei Geral de Proteção de Dados Pessoais (BRASIL, 2018), são informações que merecem cuidados adicionais, pois, associadas a outras informações, podem levar à descoberta de dados sensíveis. As instituições financeiras têm, ainda, o dever de manter o sigilo de operações financeiras de seus clientes, conforme estabelecido pela Lei Complementar nº 105/2001 (BRASIL, 2001).

Desta forma, foi implementado um tratamento para remover e anonimizar essas informações dinâmicas, priorizando a preservação da confidencialidade e proteção dos dados dos clientes, bem como o cumprimento das legislações vigentes. Além de mitigar riscos de exposição indevida, o procedimento permitiu a padronização das mensagens de erro, garantindo que todos os erros com a mesma falha conceitual fossem agrupados corretamente, independentemente dos dados variáveis, assegurando a contagem precisa das ocorrências.

3.3.2 Critérios de seleção e filtragem dos dados

O conjunto de dados brutos continha o registro de todos os tipos de erros nas transações de envio do Pix. No entanto, para permitir uma análise mais aprofundada, este trabalho concentrou-se em um subconjunto específico de dados. O critério de seleção foi o tipo de erro que, em combinação com um canal e posição do erro no programa específico, apresentou o maior volume de ocorrências de erro no período. Assim, um erro referente à extrapolação do limite de transferência via Pix foi selecionado para o estudo.

3.3.3 Criação da série temporal

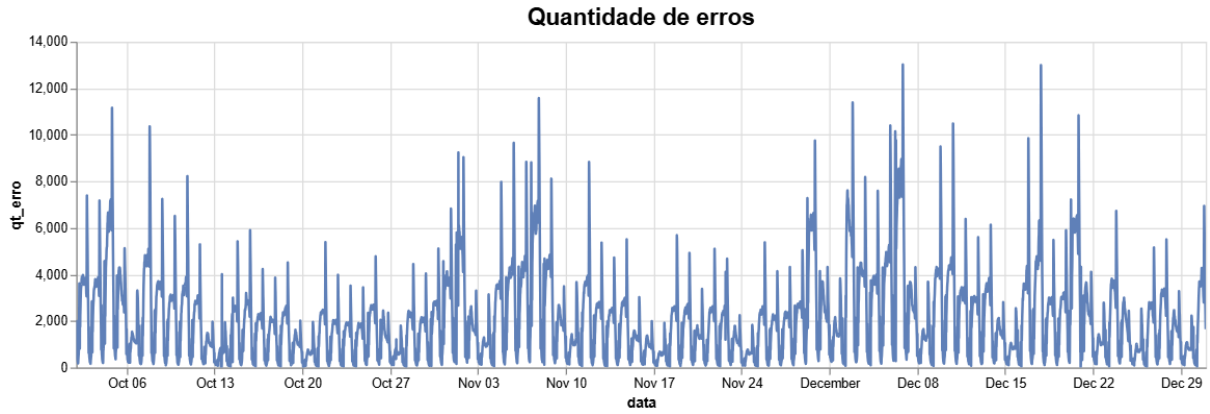
O primeiro passo foi construir um índice temporal preciso para cada evento de erro. O conjunto de dados original fornecia a data e a hora em colunas separadas: DT_MVT_PREV e HR_MVT_PREV. Estas duas colunas foram combinadas para gerar um único campo de *timestamp*, que foi convertido para um formato *datetime* e definido como índice do conjunto de dados.

Em seguida, foi aplicado um processo de reamostragem (*resampling*) para agregar os dados em uma frequência fixa de 15 minutos. Utilizando a biblioteca Pandas, foi aplicada uma função de soma (.sum()) sobre os dados reamostrados. O resultado final foi uma série temporal univariada pronta para ser utilizada nas etapas de decomposição e modelagem.

3.3.4 Análise exploratória

Após a criação da série temporal, plotamos os dados da série completa conforme apresentado na Figura 5. No gráfico foi possível observar um comportamento sazonal, com padrões diários e semanais, confirmados na etapa a seguir. Verificamos também que a amplitude das oscilações sazonais permanece relativamente constante ao longo de todo o período, mesmo com variações no nível de base da série, confirmando a característica de uma série aditiva (HYNDMAN; ATHANASOPOULOS, 2018).

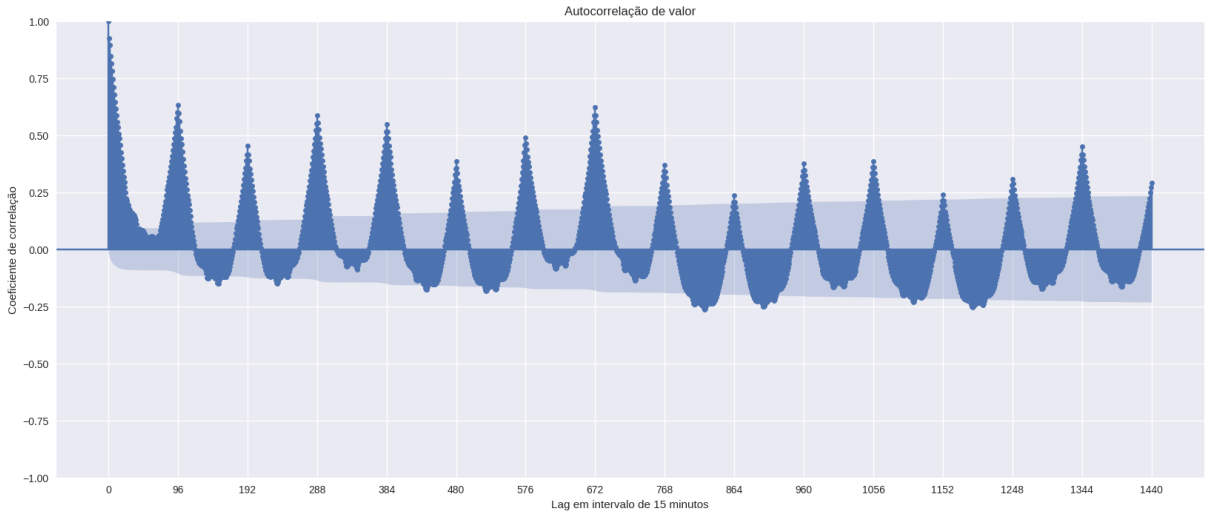
Figura 5 – Série temporal da quantidade de erros (outubro a dezembro/2024)



Fonte: Elaborado pela autora (2025).

Para confirmar a presença da sazonalidade observada visualmente, foi analisada a Função de Autocorrelação (ACF) da série, apresentada na Figura 6. A autocorrelação mede a semelhança entre uma série temporal e uma versão de si mesma defasada no tempo, plotando os coeficientes de correlação para diferentes defasagens (*lags*), permitindo identificar a presença de padrões repetitivos como a sazonalidade (HYNDMAN; ATHANASOPOULOS, 2018).

Figura 6 – Gráfico de Autocorrelação (ACF) da série temporal de erros



Fonte: Elaborado pela autora (2025).

A análise do gráfico revela picos significativos e recorrentes nos *lags* 96, 192, e assim por diante. Considerando que a série possui uma frequência de 15 minutos, o *lag* 96 corresponde a um período de 24 horas (96 * 15 minutos = 1.440 minutos = 24 horas). A repetição desses picos em múltiplos de 96 confirma a existência de uma forte sazonalidade diária. Adicionalmente, podemos observar padrões que se repetem a cada 672 *lags* (672 * 15 minutos = 10.080 minutos = 7 dias), indicando também a presença de uma sazonalidade semanal.

Os vales de correlação negativa no gráfico ACF indicam relações inversas em determinados *lags*, geralmente na metade do período sazonal (48 em torno de um ciclo de 96). Isso revela que valores altos podem ser seguidos por valores baixos 12 horas depois, refletindo oscilações naturais do erro estudado: a extrapolação do limite de transferência do Pix tende a ser maior durante a noite, quando há redução dos valores de limite por transação no período noturno para a segurança dos usuários, conforme estabelecido por Instrução Normativa do Banco Central (BACEN, 2024b). A significância estatística dos coeficientes de autocorrelação, evidenciada pelo fato de que a maioria dos picos e vales se estende para fora do intervalo de confiança, reforça que essas correlações não são aleatórias, mas sim características intrínsecas da série.

3.4 Aplicação da decomposição STL

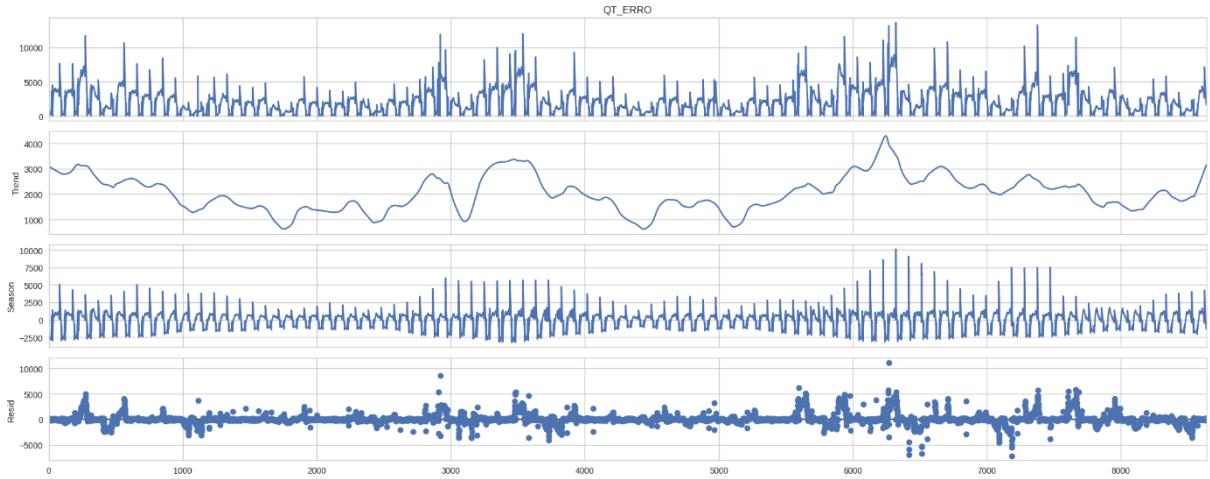
Após a confirmação da presença de componentes sazonais na série temporal de erros na fase de análise exploratória, seguimos com a aplicação da decomposição STL, fundamentada no Capítulo 2, para isolar esses padrões previsíveis. A decomposição foi implementada utilizando a função STL da biblioteca *statsmodel* da linguagem *Python*. O algoritmo opera sob um modelo de decomposição aditivo, característica observada na análise exploratória da série temporal, possibilitando o uso da função sem a necessidade de transformação logarítmica prévia dos dados (Cleveland et al., 1990).

O parâmetro de período sazonal (*period*) foi definido como 96, equivalente a um período de 24 horas, padrão identificado no item anterior. A aplicação do método *fit()* ao objeto STL resultou na separação da série temporal original em três componentes distintos descritos abaixo e ilustrados na Figura 7:

- **Tendência (*Trend*)**: O comportamento de longo prazo do número de erros.
- **Sazonalidade (*Season*)**: Os padrões repetitivos diários.
- **Resíduo (*Resid*)**: As variações não explicadas pela tendência e pela sazonalidade.

O componente de resíduo será o único produto desta etapa a ser utilizado na fase de modelagem. Esta abordagem permite que o algoritmo de detecção de anomalias se concentre exclusivamente nos desvios e ruídos não padronizados, onde as anomalias verdadeiras tendem a se manifestar, aumentando assim a precisão e a relevância das detecções.

Figura 7 – Gráfico com a decomposição STL da série temporal



Fonte: Elaborado pela autora (2025).

3.5 Treinamento do modelo *Isolation Forest*

O treinamento do modelo de detecção de anomalias foi realizado com o algoritmo *Isolation Forest*, a partir da implementação da classe *IsolationForest* da biblioteca *Scikit-learn*. O modelo utilizou como entrada exclusivamente a série temporal de resíduos, e seus hiperparâmetros foram configurados da seguinte forma:

- ***n_estimators***: Este parâmetro controla o número de árvores de isolamento na floresta. Foi definido como 100, o mesmo proposto no trabalho original de Liu, Ting e Zhou (2008).
- **max_samples**: O número de amostras a serem extraídas para treinar cada árvore foi fixado em 256, também em conformidade com o estudo de Liu, Ting e Zhou (2008).
- **contamination**: Este parâmetro informa ao modelo a proporção esperada de anomalias no conjunto de dados. Foi definido um valor inicial de 0.01 (ou 1%). Esta estimativa foi baseada no conhecimento de domínio de que anomalias são eventos raros.
- ***max_features***: Definido como 1.0, indicando que todas as *features* (neste caso, apenas uma, o próprio valor do resíduo) são consideradas para divisão em cada árvore.
- ***bootstrap***: Mantido como *False* (Falso), o que significa que as amostras para cada árvore são extraídas sem reposição.
- ***random_state***: Fixado em 0 para garantir a reprodutibilidade dos resultados. A fixação da semente aleatória assegura que, ao executar o mesmo código novamente, a mesma floresta de árvores será gerada.
- ***n_jobs***: Definido como -1, para utilizar todos os processadores disponíveis e acelerar o processo de treinamento.

Após a configuração, o modelo foi treinado com os dados de resíduo através do método *fit()*. A próxima etapa consistiu na aplicação do modelo para atribuir um *score* de anomalia e classificar cada ponto da série temporal como normal ou anômalo. Para realizar esta classificação foram usados dois métodos da classe *IsolationForest:*

1. ***.decision_function()*:** Este método foi aplicado ao conjunto de resíduos para calcular o *score* de anomalia de cada observação. Conforme descrito na fundamentação teórica, *scores* mais baixos indicam uma maior probabilidade de a observação ser uma anomalia.
2. ***.predict()***: Este método foi utilizado para obter uma classificação binária, atribuindo o rótulo -1 para as anomalias e 1 para as observações normais.

O resultado dessa etapa foi a criação de uma nova coluna no conjunto de dados, contendo a classificação de cada ponto. As observações marcadas como anômalas foram investigadas nas etapas subsequentes de análise e validação por especialistas.

4 ANÁLISE DOS RESULTADOS

4.1 Avaliação do modelo

A avaliação dos resultados de detecção de anomalias não supervisionada apresenta desafios pois, por definição, não há um gabarito de anomalias para comparação. Portanto, a validação do modelo desenvolvido neste trabalho baseou-se em análise exploratória do *score* de anomalia, inspeção visual dos pontos classificados como anômalos e validação por especialistas do domínio, com o objetivo de determinar a utilidade prática do modelo como uma ferramenta de suporte à decisão. O processo de avaliação será detalhado nos subitens a seguir.

4.1.1 Análise exploratória do s*core* de anomalia

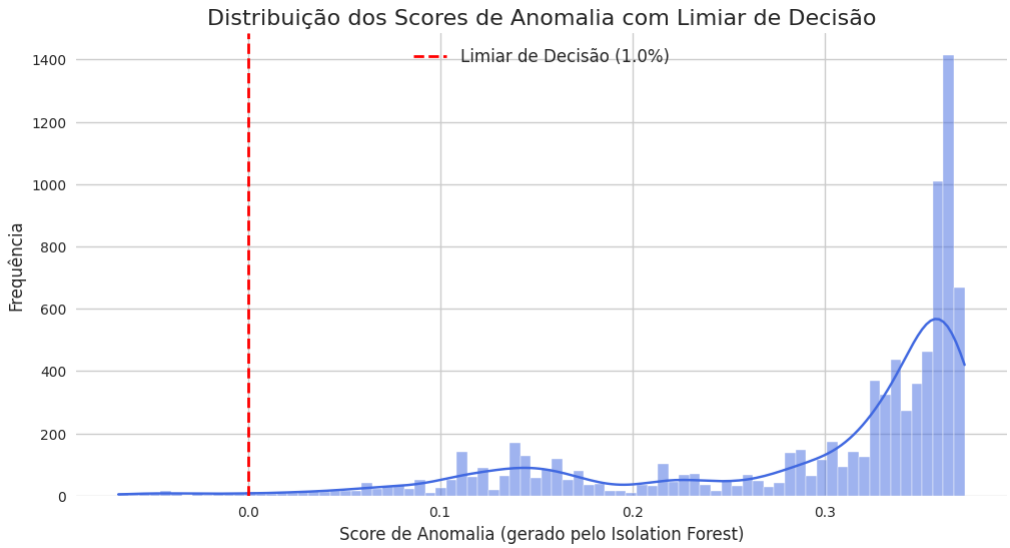
O algoritmo *Isolation Forest* não retorna apenas uma classificação binária (anomalia/normal), mas também um *score* de anomalia para cada ponto de dados. A análise visual desta distribuição permite validar se o limiar de decisão, definido implicitamente pelo hiperparâmetro *contamination*, está separando adequadamente os eventos extremos. Nesta etapa, plotamos um histograma da distribuição dos *scores* de todos os pontos da série de resíduos, gerados pelo método *.decision_function*(), conforme a Figura 8.

Para compreender a métrica apresentada no eixo X do gráfico, é fundamental entender como a função .decision_function() da biblioteca scikit-learn, utilizada neste trabalho, processa o *score* de anomalia. Primeiramente, é calculado um *offset*, que corresponde ao percentil da pontuação de anomalia original com base no valor de *contamination* definido. No caso deste estudo, com *contamination* de 1%, o *offset* será o percentil 0.99 dos *scores*. Os *scores* finais apresentados no gráfico são então calculados subtraindo-se o *offset* do *score* original. Essa transformação resulta em uma escala em que pontos mais baixos (mais negativos) indicam uma maior probabilidade de o ponto ser uma anomalia, enquanto valores mais altos (próximos de zero ou positivos) correspondem a pontos considerados normais. Essa abordagem inverte a lógica da fundamentação teórica original do *Isolation Forest*, mas facilita a interpretação visual, posicionando o limiar de decisão (o *offset* ajustado) geralmente em torno de zero (SCIKIT-LEARN, 2025).

Com essa transformação, foi possível observar no gráfico a maior concentração nos pontos considerados normais, com *scores* positivos, e uma pequena fração de observações com *scores* negativos. Esta distribuição está conforme o esperado para o algoritmo, que atribui

scores similares aos pontos normais e *scores* distintos e mais baixos às anomalias, que são mais fáceis de isolar (LIU; TING; ZHOU, 2008). A observação confirmou a premissa de que as anomalias são raras e distintas do comportamento padrão dos dados.

Figura 8 – Distribuição dos *scores* de anomalia

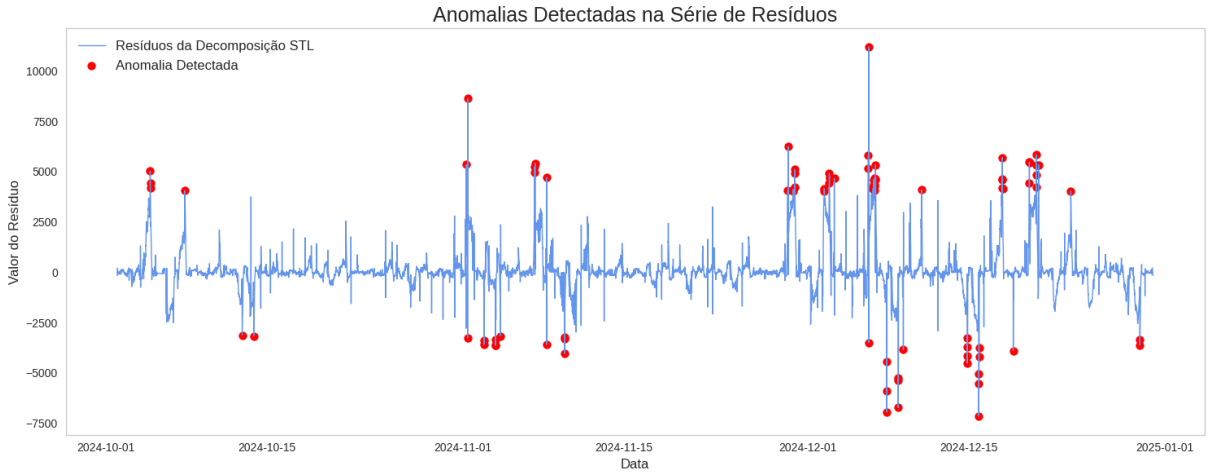


Fonte: Elaborado pela autora (2025).

4.1.2 Inspeção visual e análise de eventos

A aplicação do modelo *Isolation Forest* sobre a série de resíduos resultou na identificação de um conjunto de pontos anômalos. A Figura 9 apresenta as anomalias detectadas sobre a série de resíduos da decomposição STL.

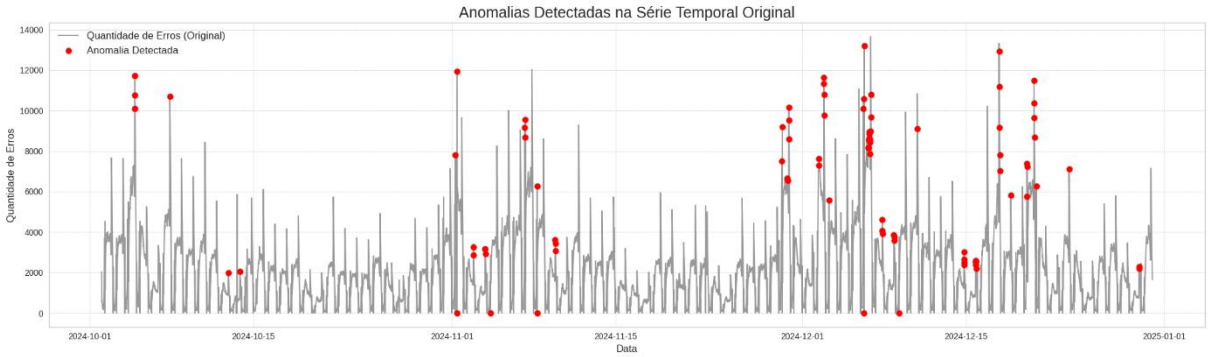
Figura 9 – Gráfico com as anomalias sobre a série de resíduos



Fonte: Elaborado pela autora (2025).

A análise do gráfico demonstra a eficácia do modelo em identificar os pontos que mais se desviam do comportamento padrão do ruído da série (cuja média é zero), mostrando que o modelo se mostrou sensível em ambas as direções. Os pontos onde o valor do resíduo é extremamente positivo indicam que a quantidade de erros observada foi significativamente maior do que esperado pelo modelo sazonal. Pontos onde o valor do resíduo é extremamente negativo indicam que a quantidade de erros foi significativamente menor do que o esperado. A Figura 10 contextualiza essas detecções na série temporal original, confirmando que os picos positivos nos resíduos correspondem aos maiores volumes de erros, enquanto que os vales negativos representam quedas inesperadas.

Figura 10 – Gráfico com as anomalias sobre a série temporal original



Fonte: Elaborado pela autora (2025).

A capacidade do modelo de detectar ambos os tipos de evento o torna uma ferramenta de monitoramento mais completa do que um sistema que define um limiar simples. A investigação junto aos especialistas, detalhada a seguir, buscou determinar as causas e o impacto de negócio tanto dos picos quanto das quedas inesperadas de erros.

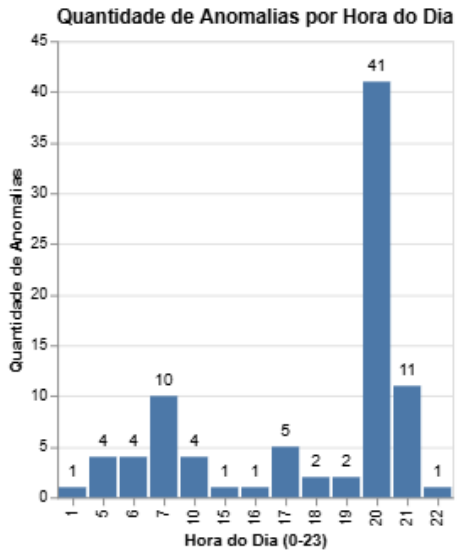
4.1.3 Análise contextual e validação por especialistas

Após a análise técnica dos resultados, a etapa final da avaliação consistiu na validação qualitativa das anomalias detectadas. Foram realizadas sessões de análise com especialistas da instituição financeira, envolvendo analistas de TI e o Gerente de TI. Primeiramente foi analisada a distribuição das anomalias por hora, a fim de observar a ocorrência de concentração em determinados períodos do dia. Depois, a validação focou-se nos dias com maior ocorrência de anomalias: 29 de novembro e 6 de dezembro de 2024, conforme detalhado a seguir.

4.1.3.1 Distribuição das anomalias por hora

No período analisado, do dia 2 de outubro a 30 de dezembro de 2024, o modelo detectou 87 pontos anômalos. A Figura 11 mostra a distribuição das anomalias por hora do dia.

Figura 11 – Distribuição das anomalias por hora do dia

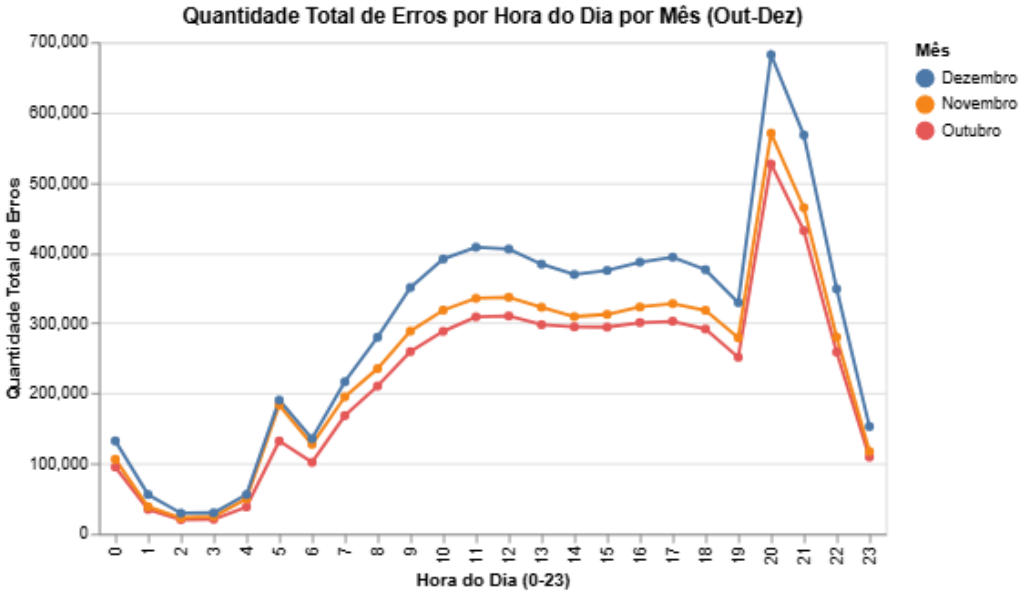


Fonte: Elaborado pela autora (2025).

O gráfico mostra que as anomalias detectadas se concentram, principalmente, no período entre 20h e 21h, totalizando 52 ocorrências, e às 7h, com 10 pontos anômalos. Iniciamos a análise verificando a quantidade de erros observados no período das 7h, nos dias em que foram detectadas anomalias. A investigação junto com analistas de TI revelou que existe um gargalo no processo de extração de dados: todos os dias há lentidão na atualização neste horário – a coleta de dados ocorre normalmente a cada 15 minutos e, por volta das 6h35, fica paralisada, retornando apenas próximo das 7h15. Este atraso de aproximadamente 40 minutos faz com que a quantidade de erros fique registrada como zero no período em que a atualização está parada, impactando as coletas posteriores que acumulam erros de um período maior, e consequentemente, têm maior probabilidade de serem detectadas como anomalia.

Ao analisarmos as anomalias detectadas no período da noite, entre 20h e 21h, junto aos especialistas da área, foi observado que o horário coincide com o início do período em que os limites noturnos para transferência via Pix entram em vigor: das oito horas da noite às seis horas da manhã. Essa redução no valor do limite permitido para transferência via Pix impacta diretamente a ocorrência de erros – podemos verificar que, neste horário, há um pico na quantidade de erros referente à extrapolação do valor permitido para o horário, conforme a Figura 12.

Figura 12 – Quantidade de erros por hora do dia por mês



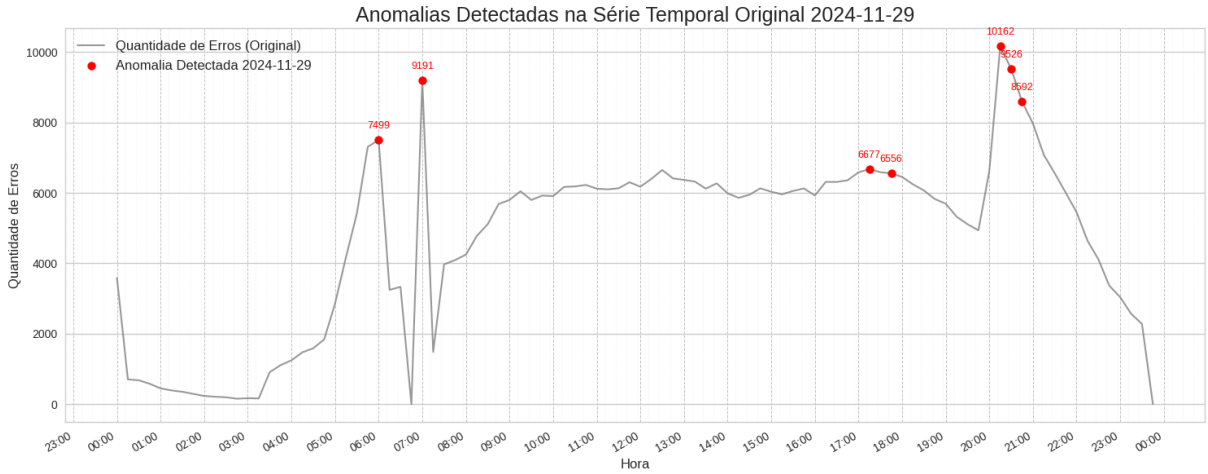
Fonte: Elaborado pela autora (2025).

A seguir, foi feita a análise dos dias com maior ocorrência de anomalias detectadas, dias 29 de novembro e 6 de dezembro de 2024, onde observaremos como os fatores analisados neste item, juntamente com outros aspectos, impactaram a detecção de anomalias.

4.1.3.2 Análise detalhada dos dias com maiores ocorrências de anomalias

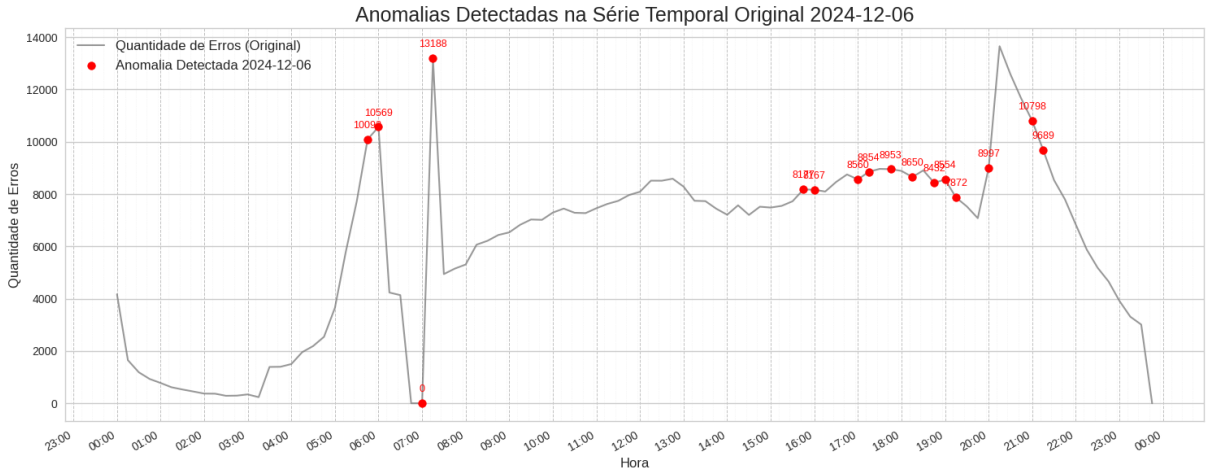
Os dias 29 de novembro e 6 de dezembro de 2024 foram marcados por uma significativa ocorrência de anomalias, com sete e dezesseis detecções, respectivamente (Figuras 13 e 14). Ambos os dias exibiram padrões de anomalias concentrados em horários similares: início da manhã (6h - 7h), meio/fim da tarde (16h - 17h) e início da noite (a partir das 20h). A validação com especialistas revelou que estes padrões foram impulsionados por uma combinação de fatores operacionais e contextuais, que serão detalhados nas subseções a seguir, para cada data.

Figura 13 – Anomalias detectadas em 29 de novembro de 2024



Fonte: Elaborado pela autora (2025).

Figura 14 – Anomalias detectadas em 6 de dezembro de 2024



Fonte: Elaborado pela autora (2025).

4.1.3.2.1 Análise do dia 29 de novembro de 2024

Neste dia, as sete anomalias detectadas foram fortemente influenciadas pela *Black Friday*, que impulsionou o volume de transações Pix a 15,8 milhões na instituição e 239,9 milhões no Brasil. A interação desse alto volume com os limites noturnos do Pix, especialmente próximo às 6h e a partir das 20h, gerou desvios significativos. Adicionalmente, a falha no fluxo de coleta de dados às 6h45, detalhada na Seção 4.1.3.1, resultou na anomalia das 7h, demonstrando a capacidade do modelo de identificar problemas na telemetria. Os pontos anômalos do fim da tarde (17h15 e 17h45) também foram impactados pelo volume excepcional de transações da *Black Friday*.

4.1.3.2.2 Análise do dia 6 de dezembro de 2024

Com dezesseis anomalias, este dia apresentou o maior número de detecções. O volume de transações foi recorde para a instituição, alcançando 17 milhões e o segundo maior volume nacional, com 250,6 milhões de operações. Este movimento foi impulsionado por fatores como a coincidência com o 5º dia útil – data de pagamento de salários e do recebimento da segunda parcela do 13º salário, além da extensão das promoções pós *Black Friday* e pós *Cyber Monday* (ação que acontece na segunda-feira pós *Black Friday*, com promoções focadas em produtos eletrônicos e de tecnologia vendidos *online*). As falhas na coleta de dados às 6h45 e 7h, explicadas na Seção 4.1.3.1, também contribuíram para as anomalias. A análise dos especialistas confirmou que o modelo capturou a complexa interação entre o comportamento do usuário, regras de negócio e condições de estresse, além de falhas no sistema de telemetria.

As descobertas apresentadas neste capítulo não apenas validam a metodologia proposta, mas também ressaltam a complexidade de interpretar eventos em um ambiente dinâmico e de alta criticidade. Com base nesses resultados e nas reflexões geradas, o próximo capítulo consolida as conclusões gerais deste estudo, destaca suas principais contribuições e aponta direções promissoras para futuros trabalhos de pesquisa e desenvolvimento.

5 CONCLUSÃO E TRABALHOS FUTUROS

5.1 Conclusão do estudo

Este trabalho investigou a aplicação de técnicas de análise de séries temporais e detecção de anomalias para identificar padrões incomuns em erros de transações do sistema Pix. Através da combinação da decomposição STL (*Seasonal-Trend decomposition using Loess*) e do algoritmo *Isolation Forest*, foi desenvolvida uma metodologia capaz de segmentar os componentes de tendência, sazonalidade e resíduos da série temporal de erros e, subsequentemente, detectar anomalias nos resíduos, que representam os desvios não explicados pelos padrões normais.

Os resultados demonstraram a eficácia da abordagem proposta na identificação de eventos anômalos, mesmo em cenários complexos influenciados por fatores externos e operacionais. A análise detalhada de dias com alta ocorrência de anomalias, como 29 de novembro (*Black Friday*) e 6 de dezembro (5º dia útil e pagamento de 13º salário), revelou que o modelo é capaz de capturar tanto desvios relacionados a picos de volume de transações e mudanças em regras de negócio (limites noturnos do Pix), quanto falhas no próprio sistema de telemetria. A validação por especialistas da instituição financeira foi fundamental para contextualizar e interpretar essas anomalias, diferenciando entre falhas críticas, variações operacionais esperadas e falsos positivos, e reforçando a importância da inteligência humana no processo de monitoramento.

Em suma, a metodologia proposta oferece uma ferramenta valiosa para a gestão proativa da qualidade e disponibilidade do serviço Pix, permitindo a identificação precoce de potenciais problemas e o aprimoramento contínuo da infraestrutura de monitoramento. As conclusões aqui apresentadas abrem caminho para as principais contribuições deste estudo, detalhadas na próxima seção

5.2 Contribuições do trabalho

As principais contribuições deste trabalho incluem:

- **Metodologia integrada:** proposição e validação de uma metodologia integrada que combina a decomposição STL para séries temporais com o *Isolation Forest* para

detecção de anomalias, especificamente adaptada para dados de erros de transações financeiras.

- **Análise contextual aprofundada:** demonstração da importância da contextualização de anomalias com eventos de negócio e operacionais, como a *Black Friday* e os limites noturnos do Pix, para uma interpretação mais precisa dos resultados.
- **Validação por especialistas:** reforço do papel fundamental da validação humana por especialistas na diferenciação entre anomalias reais e falsos positivos e na geração de *feedback* para o refinamento do modelo.
- **Detecção de falhas de telemetria:** evidenciação da capacidade do modelo de atuar como um mecanismo de *data observability*, identificando falhas no próprio sistema de coleta de dados, o que é um diferencial para a confiabilidade da infraestrutura de monitoramento.

Apesar das contribuições significativas, é fundamental reconhecer as limitações inerentes a este estudo, que serão discutidas a seguir.

5.3 Limitações

Este estudo apresenta algumas limitações que devem ser consideradas:

- **Dados específicos:** A análise foi realizada com um conjunto de dados de um único tipo de erro, em um canal e posição do erro no programa específicos, o que pode limitar a generalização dos resultados para outros tipos de erros com diferentes perfis e contextos.
- **Período de análise:** O período de análise (outubro a dezembro de 2024) é relativamente curto, o que pode não capturar todas as sazonalidades de longo prazo ou eventos anômalos raros.
- **Dependência de parâmetros:** A performance do *Isolation Fore*st é sensível à escolha de parâmetros, como o *contamination*. Embora definido com base em premissas e validação empírica neste estudo, a otimização desses parâmetros em ambientes dinâmicos, como o do Pix, é um desafio contínuo. A constante evolução dos padrões de transação e o surgimento de novos comportamentos podem exigir reajustes frequentes, tornando a manutenção da performance ideal um processo complexo e que demanda atenção constante.
- **Validação manual**: A validação por especialistas, embora crucial, é um processo manual e intensivo em recursos, o que pode ser um desafio em ambientes de alta escala.

As limitações identificadas neste trabalho servem como base para a proposição de futuras pesquisas, visando expandir e aprimorar a metodologia apresentada.

5.4 Trabalhos futuros

Com base nas conclusões e limitações deste estudo, sugere-se os seguintes trabalhos futuros:

- **Comparação com outros algoritmos:** Realizar uma comparação aprofundada com outros algoritmos de detecção de anomalias (por exemplo, *One-Class SVM, Autoencoders*) para avaliar a performance relativa em diferentes cenários de erros do Pix.
- **Aplicação em outros contextos:** Estender a metodologia para a detecção de anomalias em outros canais, tipos de erros do Pix e demais transações financeiras, validando sua generalização e adaptabilidade.
- **Análise dos dados em tempo real:** adaptar a metodologia para avaliação dos erros em tempo real, com disponibilização de painel de monitoramento com as anomalias detectadas e sistema de alerta para os especialistas em caso de detecção de anomalias.
- **Incorporação de novos atributos:** Avaliar a inclusão de informações adicionais, como dados de volume de transações efetivadas, informações de eventos externos (notícias, feriados) e informações sobre implementações de novas regras de negócio, para enriquecer o contexto da detecção de anomalias. Esses atributos poderiam permitir ao modelo diferenciar com maior precisão entre desvios estatísticos que são variações normais do sistema (por exemplo, alto volume na *Black Friday*) e anomalias que indicam problemas reais. A inclusão desses dados pode levar a uma redução de falsos positivos e a um aumento da capacidade preditiva do modelo, tornando o sistema de detecção mais inteligente e menos propenso a alarmes desnecessários.

5.5 Lições aprendidas

Antes da escolha da metodologia combinando a Decomposição STL com o *Isolation Forest*, foram realizadas outras tentativas que, embora mal-sucedidas, forneceram *insights* valiosos sobre as limitações para detecção de anomalias em séries temporais. Todas as experimentações utilizaram o mesmo conjunto de dados descrito no item 3.2 – Descrição do

conjunto de dados, e o treinamento do *Isolation Forest* seguiu a mesma configuração apresentada no capítulo 3.5 – Treinamento do modelo *Isolation Forest*. A seguir, descrevemos essas tentativas, os resultados observados e as lições extraídas de cada uma delas.

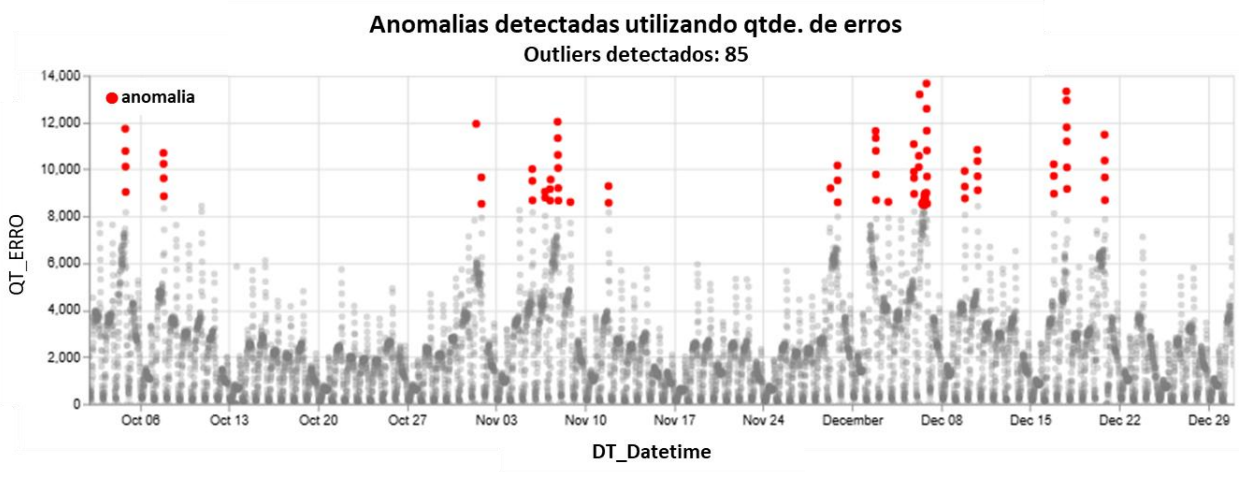
5.5.1 Aplicação direta do *Isolation Forest* na quantidade de erros

A primeira abordagem consistiu na aplicação do algoritmo *Isolation Forest* utilizando exclusivamente a *feature* "quantidade de erros" (QT_ERRO). Essa estratégia representa a forma mais simples de detecção de anomalias univariadas, na qual o modelo busca identificar valores que se desviam significativamente da distribuição observada.

Ao realizarmos a inspeção visual dos resultados, conforme ilustrado na Figura 15, observamos que o modelo apresentou baixa sensibilidade, identificando apenas os valores extremos como anomalias. Essa limitação decorre do fato de que o *Isolation Forest*, quando aplicado diretamente a uma série temporal, não considera a estrutura temporal dos dados, tratando cada observação como independente. Consequentemente, o algoritmo não consegue distinguir entre valores extremos que representam anomalias reais e aqueles que fazem parte de padrões sazonais ou tendências naturais da série.

A principal lição dessa tentativa foi a constatação de que a abordagem não foi suficiente para capturar a complexidade da série temporal, com padrões sazonais e tendências. Essa observação motivou a busca por estratégias que incorporassem informações contextuais sobre a dimensão temporal dos dados.

Figura 15 – Anomalias detectadas utilizando quantidade de erros



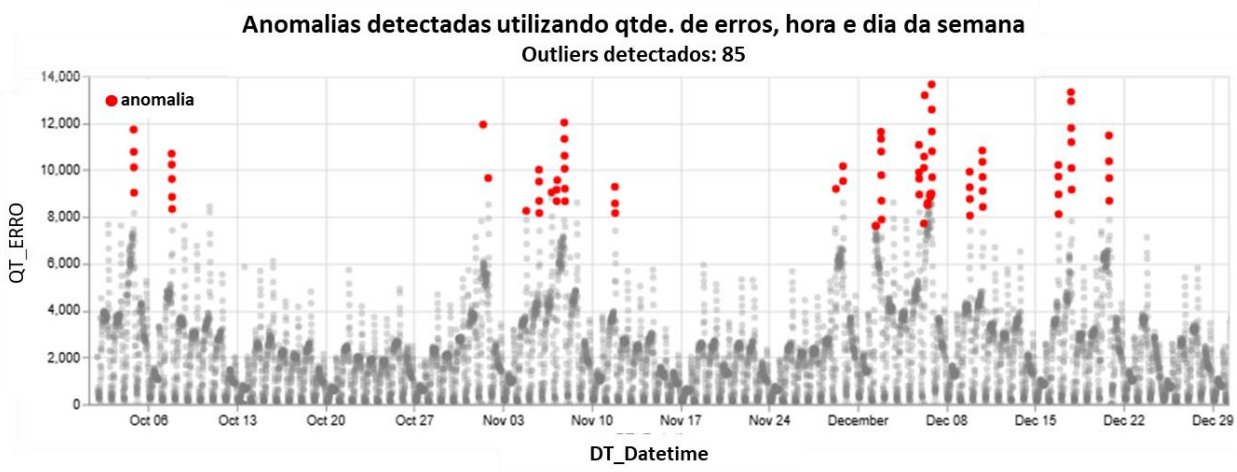
5.5.2 Criação de *features* temporais

Considerando os resultados da análise de autocorrelação, que apontaram um forte padrão semanal e a cada 24 horas (conforme descrito no capítulo 3.3.4 – Análise exploratória), foram criados os atributos "hora" e "dia da semana" a partir do campo "DT_MVT_PREV", que continha a data da coleta dos dados. O objetivo era fornecer ao modelo informações explícitas sobre os padrões de sazonalidade da série, permitindo que o *Isolation Forest* diferenciasse entre variações esperadas (relacionadas ao horário e ao dia da semana) e anomalias genuínas.

Com essa configuração, o algoritmo foi aplicado utilizando três features: "quantidade de erros", "hora" e "dia da semana". Os resultados, apresentados na Figura 16, mostraram resultados semelhantes à tentativa anterior, com tendência de concentrar as detecções em valores extremos, sem capturar adequadamente anomalias mais sutis que poderiam indicar problemas operacionais relevantes.

Essa experiência evidenciou que, embora a inclusão de *features* temporais possa fornecer contexto adicional ao modelo, adicionar variáveis relacionadas ao tempo não é equivalente a modelar a estrutura temporal e a sazonalidade de uma série. O *Isolation Forest*, por ser baseado em árvores de decisão, não captura naturalmente a dependência sequencial dos dados nem a dinâmica temporal subjacente. Essa limitação conceitual tornou-se mais evidente nesta segunda tentativa.

Figura 16 – Anomalias detectadas utilizando quantidade de erros, hora e dia da semana



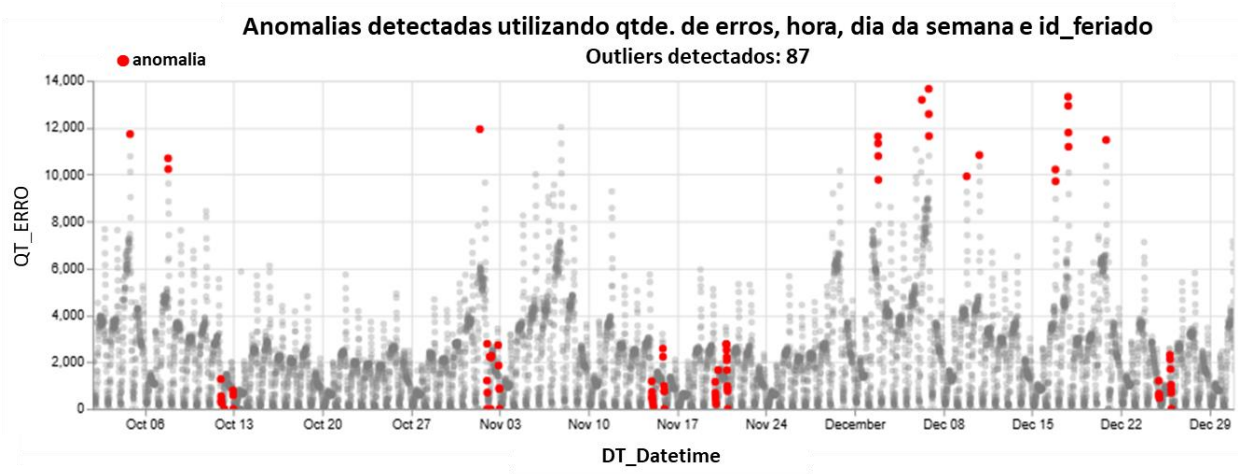
5.5.3 Inclusão da *feature* feriado

Para confirmar a hipótese de que informações sobre feriados pudessem auxiliar na detecção de anomalias, foi criado o campo binário "id_feriado", indicando o valor 1 quando a data da coleta dos dados coincidia com feriados nacionais e o valor 0 para dias úteis. A motivação para essa *feature* baseou-se na observação de que, em feriados, o volume de transações Pix tende a ser significativamente menor, o que poderia ser interpretado erroneamente como uma anomalia pelo modelo.

Essa nova *feature* foi incluída junto às variáveis analisadas anteriormente ("quantidade de erros", "hora" e "dia da semana"), e o algoritmo *Isolation Forest* foi novamente aplicado. Contrariamente ao esperado, o resultado foi insatisfatório, pois o modelo concentrou as anomalias justamente nos dias de feriado, conforme ilustrado na Figura 17. Esse comportamento indica que o algoritmo interpretou os feriados como pontos isolados na distribuição multivariada, classificando-os como anomalias em vez de reconhecê-los como padrões esperados dentro do contexto temporal.

Essa tentativa reforçou a compreensão de que o *Isolation Forest* não distingue entre valores atípicos que representam anomalias reais e aqueles que fazem parte de padrões contextuais esperados. A inclusão de *features* adicionais, sem um tratamento adequado da estrutura temporal, pode até mesmo piorar os resultados, introduzindo vieses indesejados.

Figura 17 – Anomalias detectadas utilizando quantidade de erros, hora, dia da semana e feriado



5.5.4 Síntese das lições aprendidas

As três abordagens descritas, apesar de não apresentarem resultados promissores, foram fundamentais para compreender as limitações de aplicar algoritmos de detecção de anomalias convencionais diretamente a séries temporais com padrões sazonais complexos. A principal conclusão foi que, mesmo com a inclusão de variáveis relacionadas ao tempo (como dia da semana, hora e feriados), o modelo não foi capaz de capturar adequadamente a sazonalidade da série temporal. Isso ocorre porque o algoritmo foi projetado para dados independentes e identicamente distribuídos (i.i.d.), não para dados com dependência temporal. Consequentemente, ele não modela a estrutura sequencial nem os padrões sazonais que caracterizam séries temporais. A inclusão de *features* temporais como variáveis não resolve esse problema fundamental, pois não captura a dinâmica temporal subjacente.

Essas lições motivaram a busca por alternativas metodológicas que separassem os componentes de tendência e sazonalidade dos resíduos da série, culminando na estratégia de utilizar a decomposição STL no pré-processamento, removendo os padrões previsíveis (tendência e sazonalidade), e aplicar o *Isolation Forest* exclusivamente nos resíduos.

REFERÊNCIAS

AGGARWAL, Charu C. **Outlier Analysis**. 2. ed. Cham: Springer, 2017.

AHMED, Mohiuddin; MAHMOOD, Abdun Naser; ISLAM, Md. Rafiqul. A survey of anomaly detection techniques in financial domain. **Future Generation Computer Systems**, v. 55, p. 278-288, 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167739X15000023. Acesso em: 6 mar. 2025.

BANCO CENTRAL DO BRASIL. **Diretório de Identificadores de Contas Transacionais (DICT)**. Brasília, 2020a. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/dict. Acesso em: 2 set. 2025.

BANCO CENTRAL DO BRASIL. **Estatísticas do Pix**. Brasília, 2024a. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/estatisticaspix. Acesso em: 30 ago. 2025.

BANCO CENTRAL DO BRASIL. **Instrução Normativa BCB n.º 512, de 30 de agosto de 2024**. Dispõe sobre os limites de valor para as transações no âmbito do Pix. Diário Oficial da União, Brasília, DF, 2 set. 2024b. Seção 1, p. 58.

BANCO CENTRAL DO BRASIL. **O brasileiro e sua relação com o dinheiro**. Brasília, 2024c. Disponível em: https://www.bcb.gov.br/content/cedulasemoedas/pesquisabrasileirodinheiro/Apresentacao_brasileiro_relacao_dinheiro_2024.pdf. Acesso em: 30 ago. 2025.

BANCO CENTRAL DO BRASIL. **Papel do BC**. Brasília, 2020b. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/papeldobcpix. Acesso em: 2 set. 2025.

BANCO CENTRAL DO BRASIL. **Pix**. Brasília, 2020c. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/pix. Acesso em: 30 ago. 2025.

BANCO CENTRAL DO BRASIL. **Pix vai ficar ainda mais fácil com serviço de iniciação de pagamento**. Brasília, 22 jul. 2021a. Disponível em: https://www.bcb.gov.br/detalhenoticia/568/noticia. Acesso em: 30 ago. 2025.

BANCO CENTRAL DO BRASIL. **Requisitos Técnicos para a Experiência do Usuário Pix**. Versão 7.1. Brasília, DF: BACEN, 2025a. Disponível em: https://www.bcb.gov.br/content/estabilidadefinanceira/pix/Regulamento_Pix/IV_RequisitosMinimosparaExperienciadoUsuario.pdf. Acesso em: 31 ago. 2025.

BANCO CENTRAL DO BRASIL. **Resolução BCB n.º 1, de 12 de agosto de 2020**. Institui o arranjo de pagamento Pix e aprova o seu Regulamento. Diário Oficial da União, Brasília, DF, 13 ago. 2020d. Seção 1, p. 42. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o%20BCB&numero=1. Acesso em: 31 ago. 2025.

BANCO CENTRAL DO BRASIL. **Resolução BCB n.º 177, de 22 de dezembro de 2021**. Aprova o Manual de Penalidades do Pix. Diário Oficial da União, Brasília, DF, 24 dez. 2021b. Seção 1, p. 195. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o%20BCB&numero=177. Acesso em: 31 ago. 2025.

BANCO CENTRAL DO BRASIL. **Sistema de Pagamentos Instantâneos (SPI)**. Brasília, 2025b. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/sistemapagamentosinstantaneos?ano=2025. Acesso em: 2 set. 2025.

BEYER, Betsy et al. **Site Reliability Engineering: How Google Runs Production Systems**. Sebastopol: O'Reilly Media, 2016.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 13 out. 2025.

BRASIL. **Lei Complementar nº 105, de 10 de janeiro de 2001**. Dispõe sobre o sigilo das operações de instituições financeiras e dá outras providências. Diário Oficial da União, Brasília, DF, 11 jan. 2001. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp105.htm. Acesso em: 13 out. 2025.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. **Anomaly Detection: A Survey**. ACM Computing Surveys, v. 41, n. 3, p. 1-58, 2009. DOI: 10.1145/1541880.1541882.

CLEVELAND, R. B. et al. **STL: A seasonal-trend decomposition procedure based on loess**. Journal of Official Statistics, v. 6, n. 1, p. 3-73, 1990.

COMMITTEE ON PAYMENT AND MARKET INFRASTRUCTURES (CPMI); INTERNATIONAL ORGANIZATION OF SECURITIES COMMISSIONS (IOSCO). **Principles for financial market infrastructures**. Basel: Bank for International Settlements, 2012. Disponível em: https://www.bis.org/cpmi/publ/d101a.pdf. Acesso em: 31 ago. 2025.

GLAZKOV, Artemi. **Time Series Anomaly Detection with iForest and STL**. Kaggle, 2023. Disponível em: https://www.kaggle.com/code/artemig/time-series-anomaly-detection-with-iforest-and-stl. Acesso em: 16 ago. 2025.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 2. ed. Melbourne: OTexts, 2018. Disponível em: https://otexts.com/fpp2/. Acesso em: 2 set. 2025.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. **Isolation Forest**. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 8., 2008, Pisa. Anais [...]. Pisa: IEEE, 2008. p. 413-422. DOI: 10.1109/ICDM.2008.17.

MORETTIN, Pedro A.; TOLOI, Clélia M.C. **Análise de Séries Temporais**. São Paulo: Editora Blucher, 2018. E-book. ISBN 9788521213529. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9788521213529/. Acesso em: 5 set. 2025.

SCHINDLER, Thimo F.; SCHLICHT, Simon; THOBEN, Klaus-Dieter. **Towards Benchmarking for Evaluating Machine Learning Methods in Detecting Outliers in Process Datasets**. Computers, v. 12, n. 12, p. 253, 2023. DOI: 10.3390/computers12120253.

SCIKIT LEARN. **sklearn.ensemble.IsolationForest — scikit-learn 0.23.1 documentation**, 2025. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html. Acesso em: 16 ago. 2025.