Detecção de Fraudes Financeiras na Análise de Crédito Empresarial Utilizando o Algoritmo Isolation Forest: Um Estudo Aplicado ao Setor de Bens de Consumo

João Marcelo Cattaldo Amorim
Programa de Pós Graduação
em Computação Aplicada
Universidade Presbiteriana Mackenzie
São Paulo, Brasil
joao.mca09@gmail.com

Leandro A. Silva
Programa de Pós Graduação
em Computação Aplicada
Universidade Presbiteriana Mackenzie
São Paulo, Brasil
leandroaugusto.silva@mackenzie.br

***Resumo*—As fraudes financeiras representam uma ameaça significativa ao crédito empresarial, gerando prejuízos e comprometendo a confiança nas relações comerciais. Detectar essas ocorrências é desafiador, dada a dificuldade de distinguir inadimplência severa de fraude, agravada pela escassez de dados rotulados e pela sofisticação crescente dos esquemas fraudulentos. Este artigo propõe uma metodologia baseada em aprendizado não supervisionado para detectar fraudes financeiras na concessão de crédito empresarial, com foco no algoritmo *Isolation Forest*. A abordagem contempla: preparação dos dados, filtragem por *ratings* de risco, identificação de padrões anômalos, rotulagem manual especializada e otimização do limiar de decisão com base na métrica F1-score. Os resultados indicam acurácia de 67% na detecção de inadimplência severa e fraudes, mesmo em cenários com dados desbalanceados e não rotulados. O processo sistemático de preparação dos dados contribuiu significativamente para a gestão de risco de crédito, ao destacar variáveis associadas tanto à inadimplência quanto à fraude, aprimorando a capacidade preditiva do modelo. A metodologia superou o processo tradicional conduzido por analistas, cuja taxa média de acerto é de 41%, aumentando em 39,02% a eficácia global na concessão de crédito. Como contribuição prática e teórica, recomenda-se a incorporação dessa metodologia a sistemas automatizados de decisão de crédito, promovendo a mitigação de riscos financeiros e o aprimoramento das práticas de gestão na análise de crédito empresarial.**

***Index Terms*—detecção de fraudes financeiras, concessão de crédito, Isolation Forest, aprendizado não supervisionado, detecção de anomalias**

I. Introdução

A CISP (Central de Informações São Paulo) é uma associação sem fins lucrativos, fundada em 1972, cuja missão é oferecer soluções exclusivas para a análise de risco de crédito. Segundo informações disponíveis no site institucional da CISP [10], a entidade é composta por 198 grandes indústrias de produtos de largo consumo, distribuídas em oito segmentos: Alimentos, Bebidas, Higiene Pessoal e Cosméticos, Papel, Papelaria, Utilidades Domésticas, Eletroeletrônicos e Produtos de Limpeza, representando aproximadamente 8% do Produto Interno Bruto (PIB) brasileiro.

As informações comerciais, fornecidas periodicamente pelos associados, são enriquecidas com dados provenientes de fontes públicas e privadas.

A partir dessa base com os dados enriquecidos, é gerada uma classificação de risco de *performance* (*score* de crédito), que categoriza os clientes nas faixas A e B (baixo risco), C (médio risco) e D e E (alto risco). Atualmente, cerca de 1.700 profissionais das áreas de crédito e cobrança utilizam os relatórios da CISP, acessando-os manualmente, por meio de APIs (*Application Programming Interface*) ou por motores de crédito automatizados.

Fóruns técnicos promovidos pela CISP revelaram oportunidades de aprimoramento nos modelos tradicionais de análise de risco, especialmente na identificação de comportamentos atípicos que possam indicar fraudes potenciais. Casos emblemáticos incluem crescimento anormal no volume de compras, alterações cadastrais recentes e inconsistências entre a data de fundação da empresa e seu tempo efetivo de atividade.

Dados preliminares apontam que aproximadamente 148 mil empresas — cerca de 15% da base de clientes — encontram-se em situação de inadimplência severa, acumulando dívidas superiores a R$ 2,1 bilhões,
segundo as informações internas disponibilizadas pela CISP [11].

Diante desse contexto, o objetivo deste artigo é propor e validar uma metodologia baseada em inteligência artificial, utilizando o algoritmo não supervisionado Isolation Forest para a detecção de anomalias que possam sinalizar potenciais fraudes financeiras.

Embora a magnitude das dívidas seja conhecida, ainda não é possível identificar com precisão quais desses casos configuram efetivamente fraudes financeiras.

Para garantir a confidencialidade das informações e a conformidade com a legislação vigente, bem como com as políticas e normas de segurança da informação, todos os dados utilizados neste estudo foram devidamente anonimizados por meio do algoritmo de hash SHA-256, assegurando a irreversibilidade dos identificadores e o sigilo das informações sensíveis [8].

II. Referencial Teórico

O referencial teórico apresenta as bases conceituais e metodológicas que sustentam esta pesquisa, abordando os principais conceitos, teorias e estudos relacionados ao tema.

A. Crédito Empresarial: Conceito, Concessão e Avaliação de Risco

O crédito, derivado do latim *credere* ou *creditum* (confiança), representa a capacidade de adquirir bens ou serviços com pagamento futuro, seja por pessoas ou empresas [19]. Essa prática, essencial na dinâmica econômica, envolve a disponibilização de recursos com base na promessa de quitação futura, podendo ocorrer por meio de empréstimos, financiamentos ou transações comerciais parceladas [7]. Embora existam múltiplas definições para o termo, compreender sua origem etimológica contribui para uma aplicação mais precisa e contextualizada [2].

A concessão de crédito é uma decisão estratégica e incerta, pois envolve riscos relacionados à capacidade de pagamento do tomador. Ela ocorre quando a instituição se sente segura para entregar capital ou mercadorias ao cliente [3]. Para Rossato e Pinto, trata-se de um instrumento de alavancagem de vendas, adotado por instituições financeiras, cooperativas e empresas em geral [19]. Segundo Fuhr et al., o crescimento das atividades econômicas intensifica a relevância da análise de crédito na gestão empresarial [13]. Essa relação contratual entre credor e devedor exige o desenvolvimento de metodologias robustas de avaliação de risco, baseadas em dados históricos, para mitigar a inadimplência e preservar a saúde financeira das operações [18].

Nesse contexto, a avaliação de risco e o uso de modelos de *score* de crédito tornaram-se ferramentas fundamentais. Esses modelos, baseados em informações fornecidas pelos solicitantes e em técnicas estatísticas, atribuem pontuações que diferenciam bons e maus pagadores [3]. A adoção do *score* não apenas reduz os custos operacionais e acelera a tomada de decisão, mas também promove consistência e precisão nas análises [13]. Assim, a classificação do risco de crédito assume um papel central nas instituições financeiras, sendo amplamente reconhecida como suporte técnico essencial à sustentabilidade do sistema creditício [12].

B. Fraudes Financeiras e Detecção de Fraudes

A fraude financeira representa uma ameaça crescente, com impactos negativos significativos no setor financeiro e na sociedade. Embora a mineração de dados seja eficaz na detecção de fraudes, ela enfrenta desafios devido à constante mudança nos perfis comportamentais e à similaridade entre transações fraudulentas e legítimas [16].

A fraude é um processo sistemático de ações cujo objetivo é distorcer dados intencionalmente e que se estabelece quando agentes internos e externos da companhia possuem a intenção de agir dissimuladamente [21]. A Associação dos Investigadores de Fraude Certificados (ACFE) classifica as fraudes em três categorias principais:

1) **Corrupção**: Envolve subornos ou uso indevido de bens públicos.
2) **Apropriação indevida de ativos**: Inclui fraudes que impactam ou não diretamente o caixa da empresa.
3) **Demonstrações financeiras fraudulentas**: Compreendem receitas fictícias e ocultação de passivos e despesas.

Entre as fraudes mais comuns no Brasil, as relacionadas a contas a pagar e contas a receber representam 23% do total de investigações de fraudes conduzidas, sendo detectadas, em sua maioria, por denúncias anônimas, denúncias nominais e atividades de auditoria interna.

A prevenção de fraudes na concessão de crédito é um tema de crescente importância no contexto financeiro brasileiro, especialmente diante do aumento expressivo de tentativas fraudulentas, como roubo de identidade e apresentação de informações falsas. O cenário atual exige que instituições financeiras e empresas adotem estratégias robustas e proativas para identificar e mitigar esses riscos, protegendo a integridade do sistema financeiro e os consumidores [22].

Entre os principais desafios da prevenção de fraudes na concessão de crédito estão:

- **Risco de inadimplência**: Requer uma análise precisa da capacidade do cliente de cumprir suas obrigações financeiras, utilizando históricos de crédito e modelos de risco.
- **Avanço tecnológico**: Acompanhamento da evolução das tecnologias para garantir segurança e inovação nos processos de análise.
- **Fraudes sofisticadas**: Enfrentar técnicas cada vez mais avançadas de roubo de identidade e falsificação documental, demandando medidas rigorosas de validação.

C. Detecção de Anomalias

A detecção de anomalias é um método para identificar ocorrências suspeitas de eventos e itens de dados que podem causar problemas para as autoridades competentes [14]. As anomalias nos dados geralmente estão associadas a questões como problemas de segurança, falhas em servidores, fraudes bancárias, falhas estruturais em edifícios, defeitos clínicos, entre outros.

D. Aprendizado de Máquina e Detecção de Fraudes

O aprendizado de máquina (ML – Machine Learning), subárea da inteligência artificial (IA), estuda algoritmos capazes de melhorar seu desempenho com a experiência, sem necessidade de programação explícita [16]. Esses algoritmos constroem modelos a partir de dados de treinamento, permitindo previsões ou decisões automatizadas.

Instituições financeiras adotam ML em diversas aplicações, como análise de risco, previsão de falências, cotações, segmentação de mercado e, principalmente, detecção de fraudes [16]. Nessa área, o objetivo é identificar transações atípicas em contextos diversos (financeiro, comercial, contábil), distinguindo eventos legítimos de fraudulentos por meio de modelos de classificação binária [14].

O ML é eficaz na extração de padrões em grandes volumes de dados, aumentando a precisão nas decisões. No entanto, enfrenta desafios como a anonimização dos dados e o desbalanceamento entre classes, que dificultam a reprodução de estudos e afetam o desempenho dos modelos [4].

E. Algoritmo Isolation Forest

O Isolation Forest é definido como um algoritmo de detecção de anomalias eficiente e independente de modelos, capaz de identificar amostras anômalas sem a necessidade de construir perfis detalhados dos dados [9]. O algoritmo utiliza árvores para isolar instâncias anômalas, baseando-se no comprimento médio do caminho entre os nós analisados e o nó raiz. Essa abordagem elimina a necessidade de cálculos de distância e densidade, aproveitando as características das anomalias — pontos raros e distintos — para separá-las rapidamente dos dados regulares, com eficiência temporal linear e baixo consumo de memória.

O Isolation Forest como um algoritmo não supervisionado que utiliza um conjunto de árvores de decisão para identificar anomalias [14]. Em vez de calcular distâncias entre pontos ou criar perfis de instâncias normais, ele isola anomalias particionando os dados.

1) Etapas de Implementação do Isolation Forest

A implementação do Isolation Forest envolve três etapas principais: seleção de hiperparâmetros, treinamento do modelo e avaliação com métricas apropriadas.

2) Seleção de Hiperparâmetros

Os dois hiperparâmetros principais do Isolation Forest são:

- **Número de Árvores:** Define a quantidade de árvores na floresta. Mais árvores podem aumentar a precisão na detecção de anomalias, mas também elevam a complexidade computacional.
- **Nível de Contaminação:** Representa a proporção estimada de anomalias no conjunto de dados. Esse parâmetro deve ser ajustado com validação cruzada ou conhecimento especializado para equilibrar precisão e revocação.

3) Treinamento do Modelo

Após definir os hiperparâmetros, o modelo é treinado utilizando dados pré-processados [17]. Durante o treinamento, as árvores de isolamento dividem o espaço de características aleatoriamente até isolar cada ponto de dado. Pontuações de anomalia são atribuídas com base no comprimento médio do caminho necessário para isolar cada ponto, sendo os valores mais altos associados a possíveis anomalias.

4) Seleção e Impacto do Limiar de Anomalia

O limiar de pontuação de anomalia diferencia dados normais de anômalos, afetando o equilíbrio entre precisão (capacidade de evitar falsos positivos) e revocação (capacidade de detectar a maioria das anomalias). Métodos avançados, como curvas ROC (*Receiver Operating Characteristic*) e AUC (*area under the ROC curve* e métricas específicas, ajudam a selecionar limiares adequados, considerando o desbalanceamento das classes e a eficácia do modelo.

5) Avaliação do Modelo com métricas apropriadas

A avaliação do modelo considera quatro métricas principais:

1) **Precisão:** Proporção de anomalias corretamente identificadas entre todas as classificadas como ta
2) **Revocação:** Capacidade do modelo em detectar a maioria das anomalias reais.
3) **F1-Score:** Média harmônica entre precisão e revocação, ideal para classes desbalanceadas.
4) **ROC-AUC:** Mede a performance geral do modelo, considerando a relação entre verdadeiros e falsos positivos em diferentes limiares.

III. Trabalhos Relacionados

A detecção de fraudes financeiras por meio de aprendizado não supervisionado tem ganhado destaque, especialmente em cenários com dados reais e ausência de rótulos, mostrando-se eficaz na identificação de padrões anômalos em grandes volumes de dados.

Chen et al. (2023) aprimoraram o Isolation Forest com ajustes dinâmicos no parâmetro de contaminação, enquanto Brzozowska et al. (2023) exploraram a integração de técnicas como DBSCAN e Autoencoders com validação estatística [9], [6]. Soares et al. (2024) aplicaram LOF e DBSCAN em dados contábeis reais, confirmando sua eficácia na detecção de fraudes estruturais [21]. Bhati (2024) obteve F1-score superior a 60% ao utilizar Autoencoders densos em fraudes com cartões de crédito [4]. Meduri (2024), por sua vez, propôs validação cruzada automatizada no uso do Isolation Forest, aprimorando sua generalização [17].

Esses estudos reforçam a robustez das abordagens não supervisionadas e sugerem avanços como a combinação de modelos, validação cruzada estratificada e incorporação de feedback humano para aumentar a confiabilidade das análises.

IV. Estudo de Caso: Detecção de Fraude na CISP

A. Identificação de Padrões Comportamentais

Adicionalmente às soluções tecnológicas oferecidas, a CISP fomenta a interação contínua entre os profissionais de crédito e cobrança das empresas associadas, por meio de plataformas colaborativas e encontros periódicos. Esses fóruns têm-se mostrado ambientes frutíferos para o compartilhamento de experiências práticas e a identificação de padrões comportamentais associados a potenciais clientes fraudulentos.

Em diversas ocasiões, foram relatados casos de empresas que, apesar de apresentarem inicialmente boas classificações de risco de crédito, passaram a exibir, em um curto intervalo de tempo, comportamentos típicos de inadimplência severa.

A partir dessas discussões, foram levantadas hipóteses sobre padrões de comportamento. Entre os principais padrões observados, destacam-se:

1) **Alterações Cadastrais**: Modificações em dados cadastrais relevantes, como razão social, endereço, CNAE (Classificação Nacional de Atividades Econômicas) e situação cadastral da empresa.
2) **Deterioração do Score de Crédito**: Redução repentina do score de performance de crédito, com transição de faixas de baixo risco (A ou B) para alto risco (D ou E).
3) **Inconsistências entre Tempo de Atividade e Data de Fundação**: Divergências entre a data de constituição da empresa e registros de atividades recentes.
4) **Aumento Anormal no Volume de Consultas**: Picos incomuns no número de consultas realizadas sobre o CNPJ da empresa em um curto espaço de tempo, o que pode indicar movimentações suspeitas no mercado.
5) **Crescimento Atípico no Volume de Compras**: Ex-

pansões abruptas e não justificadas no volume de compras efetuadas pela empresa.

6) **Crescimento Atípico no Débito Total**: Elevações súbitas no valor total de débitos.
7) **Aumento no Número de Associadas com Débitos Relacionados**: Ampliação significativa no número de empresas associadas impactadas por inadimplência do mesmo cliente.
8) **Crescimento Atípico no Valor do Maior Acúmulo de Débitos**: Incremento expressivo no maior valor isolado.
9) **Aumento no Número de Associadas Relacionadas ao Maior Acúmulo**: Expansão do número de empresas afetadas por esse maior acúmulo, reforçando a hipótese de impacto coletivo.

V. Materiais e Métodos

O estudo foi desenvolvido na plataforma gratuita *Google Colab*, utilizando back-end do *Google Compute Engine* com interpretador *Python 3*, 12,7 GB de RAM e 225,8 GB de armazenamento temporário.

A implementação foi realizada integralmente em *Python*, com apoio das bibliotecas: Pandas (manipulação de dados), NumPy (operações numéricas), Hashlib (anonimização de CNPJs), Matplotlib e Seaborn (visualização gráfica), e Scikit-learn (execução do algoritmo *Isolation Forest*).

A. Conjunto de Dados utilizados na Pesquisa

O conjunto de dados utilizado nesta pesquisa é composto por 2.191.293 observações e 19 variáveis. A base contempla 647.132 CNPJs distintos, refletindo uma ampla representatividade do universo empresarial analisado. Importa destacar que não foram identificadas duplicatas.

Nome da Variável	Tipo da Variável	Descrição	Exemplo de Valor
customerID	Categórica	CNPJ Identificação única da empresa (Cadastro Nacional da Pessoa Jurídica)	43454446000107
updated	Temporal	Data da informação (Mês e Ano) formato YYYY-MM	2025-05
rating	Categórica	Classificação do risco de performance empresa (score de crédito)	A
name	Categórica	Nome oficial da empresa	CISP Central de Informações São Paulo
cnae	Categórica	CNAE (Código da Atividade Econômica Principal da Empresa)	94.11-1-00
status	Categórica	Situação cadastral da empresa junto a Receita Federal do Brasil	ATIVA
registration	Temporal	Data de cadastro da empresa na CISP	1972-11-30
foundation	Temporal	Data de Fundação da empresa	2000-01-15
address	Categórica	Endereço da empresa	R DO BOSQUE 1589 ANEXO II ANDAR 15
neighborhood	Categórica	Bairro da empresa	BARRA FUNDA
city	Categórica	Cidade da empresa	SAO PAULO
state	Categórica	Estado da empresa (Unidade Federativa)	SP
zipCode	Numérica	CEP da empresa	1136001
queries	Numérica	Quantidade de consultas realizadas pelos associados para o CNPJ	33
amountGreaterAccumulation	Numérica	Valor do Maior Acúmulo (Valor consolidado contendo todos associados)	500000,00
quantityGreaterAccumulation	Numérica	Quantidade de associadas que compõem o Valor o Maior Acúmulo	10
amountDebit	Numérica	Valor do Débito Total (Valor consolidado contendo todos associados)	300000,00
quantityDebit	Numérica	Quantidade de associadas que compõem o Débito Total	5
lastPurchase	Temporal	Data da Última Compra (mais recente)	2024-12-30

Figura 1. Dicionário de Dados Original

A Figura 1 apresenta o dicionário de dados no qual são detalhadas as variáveis analisadas.

B. Criação de Novas Variáveis

Com o objetivo de avaliar indícios de fraudes financeiras, foram criadas novas variáveis categorizadas com base em hipóteses.

1) Validação da Situação Cadastral na Receita Federal do Brasil

A variável `status` representa a situação cadastral do CNPJ na Receita Federal do Brasil, podendo assumir os seguintes valores: *Ativa*, *Baixada*, *Inapta*, *Suspensa* ou *Nula*. Para fins de avaliação de risco na concessão de crédito, a identificação de CNPJs com status divergente de 'Ativa' é essencial.

A fim de facilitar esse processo, foi criada a variável booleana `statusActive`, definida como `True` para registros com status igual a "Ativa" e `False` nos demais casos.

2) Identificação da Existência de Alterações no Endereço ou na Razão Social

Com o intuito de validar a hipótese de existência de alterações cadastrais relevantes ao longo do tempo, foram criadas variáveis booleanas — `addressChanged` e `nameChanged` — destinadas a sinalizar modificações nas respectivas informações.

A variável `addressChanged` está relacionada às alterações nos campos de endereço, abrangendo as colunas: `address`, `neighborhood`, `city`, `state` e `zipCode`. Por sua vez, a variável `nameChanged` considera modificações exclusivamente na coluna `name`.

O processo de identificação das alterações foi conduzido por meio das seguintes etapas:

1) Ordenação dos registros com base nas colunas `customerID` e `updated`, permitindo a análise cronológica dos dados;
2) Comparação entre os valores atual e anterior das variáveis mencionadas, considerando registros pertencentes ao mesmo `customerID`;
3) Detecção de modificações específicas no histórico cadastral de cada cliente;
4) Armazenamento dos resultados das comparações em colunas booleanas, sinalizando a ocorrência ou não de alterações.

3) Identificação da Existência de Alterações no CNAE

Com o intuito de validar a hipótese de existência de alterações no CNAE (Classificação Nacional de Atividades Econômicas), foi criada a variável booleana `cnaeChanged`, destinada a sinalizar modificações nas respectivas informações associadas a cada entidade jurídica, uma vez que mudanças frequentes ou recentes na atividade econômica podem indicar tentativas de ocultação de risco, reestruturação societária suspeita ou mesmo a preparação para práticas fraudulentas.

4) Definição do Nível de Risco com Base na Fundação da Empresa

Com o intuito de avaliar o risco da empresa com base na sua data de fundação, foi criada a variável booleana `riskFoundation`, destinada a sinalizar o risco associado a empresas recém-constituídas. Sob a ótica da gestão de risco de crédito, a data de fundação é um indicador relevante, uma vez que empresas muito recentes podem representar maior vulnerabilidade, incerteza quanto à sua capacidade operacional e, potencialmente, risco de inadimplência ou fraude.

Para a criação da variável `riskFoundation`, foi elaborada a variável `yearsSinceFoundation`, que representa a quantidade de anos decorridos desde a data de fundação da empresa até a data atual. Com base nesse valor, definiu-se as seguintes faixas de risco: menos de 5 anos indica risco baixo; entre 5 e 15 anos, risco médio; e acima de 15 anos, risco alto.

Por fim, a variável `riskFoundation` foi atribuída como

`True` ou `False` conforme a verificação do risco identificado em cada caso.

5) Identificação de Empresas Fundadas Recentemente

Para avaliar se a empresa foi fundada recentemente, foi criada a variável booleana `isNewFoundation`, que recebe o valor `True` caso a empresa tenha sido fundada dentro dos últimos 3 anos, e `False` caso contrário. A identificação de empresas recém-constituídas é fundamental no contexto da gestão de risco de crédito, uma vez que organizações com pouco tempo de atividade tendem a apresentar maior vulnerabilidade financeira.

6) Identificação de Empresas com risco de inatividade comercial

Visando avaliar a existência de risco de inatividade comercial associado a empresas que, embora constituídas há mais tempo, não realizaram compras nem mantiveram relacionamento em determinado período, foi criada a variável indicadora denominada *registrationAlert*. Na perspectiva da gestão de risco de crédito, tal comportamento sugere que a empresa possui um CNPJ ativo, mas não estabelece vínculos comerciais, o que pode indicar elevado risco.

7) Identificação de aumento anormal no volume de consultas

Com a necessidade de avaliar o comportamento no volume de consultas, foi criada uma variável denominada *queriesIncreaseAlert*, destinada a identificar aumentos superiores a 30% em comparação à média móvel dos últimos 3 meses. Caso essa condição seja verificada, a variável recebe o valor *true*; caso contrário, o valor *false*. Um acréscimo de 30% no volume de consultas para uma determinada empresa pode indicar um risco elevado.

8) Identificação de Atividade Comercial Recente

Para a validação da existência de compra recente, caracterizada como aquela realizada nos últimos 3 meses, considera-se que a empresa pode apresentar inadimplência ou até mesmo estar em processo de recuperação judicial, mas ainda mantém um relacionamento ativo. Dessa forma, esse comportamento não indica, necessariamente, uma possível fraude.

Para realizar essa validação, foi criada a variável *isLastPurchase*, que compara a data atual, retroagida em 3 meses, com a data da última compra; caso a condição seja verdadeira, a variável assume o valor *true*. Caso não existam compras nos últimos 3 meses, a variável recebe o valor *false*.

9) Identificação de Crescimento Atípico no Valor do Maior Acúmulo

O valor do maior acúmulo, representado pela variável *amountGreaterAccumulation*, corresponde à maior exposição (maior saldo em aberto) que a empresa teve com o tomador de crédito nos últimos 12 meses. Em muitos casos, esse valor representa de forma mais fiel o porte da empresa do que o limite de crédito atribuído ou solicitado.

Para identificar aumentos superiores a 30% nesse valor, o que pode indicar um risco elevado, foi criada a variável *amountGreaterAccumulationAlert*. Esta variável assume o valor *true* caso seja detectado um acréscimo atípico de 30% ou mais, com base na comparação com a média móvel dos últimos 3 meses; caso contrário, assume o valor *false*.

10) Identificação de Crescimento Atípico na Quantidade de Empresas que Compõem o Maior Acúmulo

Para verificar se houve um acréscimo na quantidade de empresas que compõem o maior acúmulo nos últimos 3 meses, foi criada a variável booleana *quantityGreaterAccumulationAlert*, que recebe o valor *true* caso ocorra um aumento superior a 30% na quantidade de empresas em relação à média móvel do mesmo período; caso contrário, recebe o valor *false*. A ocorrência de um crescimento atípico, caracterizado por ultrapassar 30%, representa um forte indicativo de risco elevado.

11) Identificação de Crescimento Atípico no Débito Atual

O valor do débito atual, representado pela variável *amountDebit*, retrata a dívida que a empresa possui, resultante da soma dos títulos a vencer e vencidos calculados nos últimos 12 meses. A identificação de um crescimento atípico, caracterizado por um aumento superior a 30% em comparação com a média móvel dos últimos 3 meses, sinaliza uma deterioração preocupante na dívida da empresa.

12) Identificação de Crescimento Atípico na Quantidade de Empresas que Compõem o Débito Atual

A quantidade de empresas que compõem o valor do débito atual, representada pela variável *quantityDebitAlert*, retrata a quantidade de empresas com as quais a empresa possui dívida, resultante da soma dos títulos a vencer e vencidos, calculados nos últimos 12 meses. A identificação de um crescimento atípico, caracterizado por um aumento superior a 30% em comparação com a média móvel dos últimos 3 meses, sinaliza uma deterioração preocupante na quantidade de fornecedores que compõem a dívida.

C. Consolidação do Novo Conjunto de Dados

A Figura 2 representa o novo conjunto de dados que será utilizado para a execução das próximas etapas do processo desta pesquisa. Este dicionário de dados apresenta o nome das variáveis criadas na etapa anterior. O conjunto possui 16 colunas e um total de 2.138.375 linhas, proporcionando uma base robusta para as análises subsequentes.

Nome da Variável	Tipo da Variável	Descrição	Exemplo de Valor
customerID	Categórica	CNPJ Identificação única da empresa (Cadastro Nacional da Pessoa Jurídica)	43454446000107
updated	Temporal	Data da informação (Mês e Ano) formato YYYY-MM	2025-05
rating	Categórica	Classificação do risco de performance empresa (score de crédito)	A
statusActive	Numérica	Alteração no status da empresa (ativo ou inativo)	FALSE
addressChanged	Numérica	Alteração no endereço da empresa	FALSE
cnaeChanged	Numérica	Alteração na atividade econômica da empresa	FALSE
nameChanged	Numérica	Alteração no nome da empresa	FALSE
riskFoundation	Numérica	Risco associado à data de fundação da empresa	FALSE
isNewFoundation	Numérica	Empresa é nova (fundada nos últimos 3 meses)	FALSE
registrationAlert	Numérica	Alerta de registro recente da empresa	FALSE
queriesIncreaseAlert	Numérica	Alerta de aumento anormal nas consultas à empresa	FALSE
isLastPurchase	Numérica	Empresa realizou compras nos últimos 3 meses	FALSE
amountGreaterAccumulationAlert	Numérica	Alerta de crescimento atípico no valor que compõe o maior acumulador de crédito	FALSE
quantityGreaterAccumulationAlert	Numérica	Alerta de crescimento atípico na quantidade que compõe o maior acumulador de crédito	FALSE
amountDebitAlert	Numérica	Alerta de crescimento atípico no valor do débito atual	FALSE
quantityDebitAlert	Numérica	Alerta de crescimento atípico na quantidade que compõe o débito atual	FALSE

Figura 2. Dicionário de Dados Novo

D. Desafios Identificados na Base de Dados

A análise exploratória revelou três desafios centrais de qualidade dos dados: desbalanceamento de classes, sparsidade e ruído.

O desbalanceamento ficou evidente em variáveis críticas,

como *quantityGreaterAccumulationAlert* e *amountGreaterAccumulationAlert*, com mais de 97

A sparsidade, por sua vez, afetou variáveis como *registrationAlert* (8

Por fim, ruídos foram observados em variáveis como *riskFoundation*, cujas categorias — "Baixo", "Médio" e "Alto" risco — apresentaram sobreposições semânticas que impactam a distinção de perfis legítimos e fraudulentos.

Tais limitações exigiram estratégias de engenharia de atributos, uso de métricas robustas e modelos como o *Isolation Forest*, capazes de lidar com desbalanceamento e dados esparsos.

E. Metodologia Experimental

A Figura 3 apresenta a metodologia proposta nesta pesquisa para a detecção de fraudes financeiras no contexto da concessão de crédito.

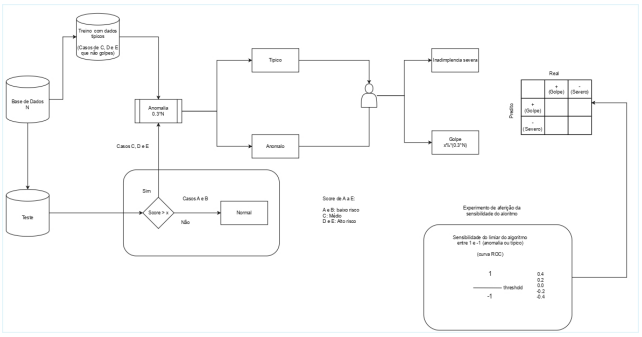

Figura 3. Metodologia para a detecção de fraudes financeiras

1) Filtragem do Conjunto de Dados

Na primeira etapa, realizou-se a filtragem do conjunto de dados, previamente elaborado de forma detalhada. Esta filtragem selecionou registros classificados com ratings C (médio risco), bem como D e E (alto risco), considerando uma linha do tempo de 12 meses, na qual empresas inadimplentes são categorizadas com esses ratings.

A metodologia proposta busca compreender o fenômeno das fraudes financeiras por meio da identificação de empresas que já se encontram inadimplentes, seja em decorrência de fraude financeira, seja por inadimplência severa.

2) Aplicação do Modelo Isolation Forest

Na segunda etapa, o processo foi complementado com a aplicação do modelo Isolation Forest, amplamente reconhecido pela sua eficácia na detecção de anomalias em grandes volumes de dados.

A implementação foi realizada inicialmente com a remoção dos casos com valores ausentes (*missing values*), sem a aplicação de técnicas de imputação. Outro ponto importante foi a utilização da técnica de *one-hot encoding* nas variáveis categóricas, de modo a atender aos pré-requisitos do modelo.

Assim, o modelo foi configurado com os seguintes parâmetros:

O conjunto de dados, previamente filtrado para conter apenas empresas com ratings C, D e E, resultou em 587.167 registros. Após a aplicação do Isolation Forest, foram identificados 29.337 registros anômalos, representando aproximadamente 5% do total, conforme o parâmetro de contaminação definido. O restante, 557.830 registros, foi classificado como normal.

Tabela I
PARÂMETROS DO MODELO ISOLATION FOREST

Parâmetro	Valor
`n_estimators`	100
`max_samples`	`'auto'`
`contamination`	0.05
`random_state`	42
Anomalias	29.337 (5%)
Não-anomalias	557.830 (95%)

3) Rotulagem Manual por Especialista

O terceiro passo consistiu na submissão de alguns casos reais para um especialista — representado por um analista financeiro — com o intuito de selecionar exemplos que representassem fielmente os casos reais, tanto de anomalias quanto de registros normais.

Durante a análise realizada pelo especialista, foi possível observar que, mesmo em casos extremos de inadimplência severa, nem sempre há correspondência com fraudes financeiras. Em diversas situações, empresas apresentavam inadimplência severa em decorrência de processos de recuperação judicial, mas ainda assim mantinham um relacionamento comercial ativo e legítimo.

Como não havia, previamente, dados rotulados disponíveis, foi necessária a atuação do especialista para realizar essa rotulação manual, possibilitando, posteriormente, a avaliação da eficácia do modelo por meio da aplicação de métricas de desempenho apropriadas.

Dada a dificuldade em rotular casos reais — por ser um processo manual e moroso — foram selecionados 6 casos reais de fraudes financeiras e 9 casos reais de inadimplência severa, que, no entanto, não configuraram fraudes.

4) Análise do Processo Manual

O cálculo da porcentagem de eficiência do processo manual de identificação de fraudes financeiras foi conduzido mediante uma análise quantitativa, considerando a quantidade de empresas que, mesmo apresentando informações suspeitas — evidenciadas na etapa de preparação dos dados —, tiveram crédito concedido.

A partir da extração dos casos reais de fraudes financeiras, constatou-se que os analistas acertaram aproximadamente **41%** dos casos, deixando de identificar o restante. Esse resultado evidencia as limitações do processo decisório atual, caracterizado por uma baixa eficácia na detecção de fraudes, especialmente em cenários marcados pela presença de sinais suspeitos que não foram devidamente considerados na tomada de decisão.

5) Avaliação de Desempenho e Busca do Melhor Limiar de Anomalia

Após a rotulagem dos casos reais, o quarto passo consistiu em submeter esses casos — tanto as fraudes quanto as não fraudes — ao modelo já treinado, com o intuito de avaliar o desempenho na distinção entre inadimplência severa e fraudes financeiras, identificando corretamente os casos onde, de fato, houve fraude.

Para a execução desta etapa, foi elaborado o Algoritmo 1, apresentado a seguir, que descreve a busca do melhor limiar (*threshold*) para a decisão.

Algoritmo 1 Busca pelo Melhor Limiar de Decisão

Entrada:
  $anomaly\_scores \leftarrow$ pontuações de anomalia do Isolation Forest
  $true\_labels \leftarrow$ rótulos reais (fraude ou não fraude)

Inicializar:
  $melhor\_limiar \leftarrow 0$ ▷ Melhor limiar encontrado
  $melhor\_f1 \leftarrow 0$ ▷ Melhor F1-score encontrado

for all limiar em range de valores possíveis **do**
  $pred\_labels \leftarrow$ classificar $anomaly\_scores$ com $limiar$
  $f1 \leftarrow$ calcular F1-score entre $pred\_labels$ e $true\_labels$
  if $f1 > melhor\_f1$ **then**
    $melhor\_f1 \leftarrow f1$
    $melhor\_limiar \leftarrow limiar$
  end if
end for

Saída:
  $melhor\_limiar$ e $melhor\_f1$

Este algoritmo realiza a varredura de diversos valores de limiar dentro de uma faixa pré-definida, utilizando como critério de otimização a maximização da métrica *F1-score*.

A aplicação deste algoritmo permitiu estabelecer um ponto de corte mais adequado para distinguir entre fraudes financeiras e inadimplências severas, proporcionando ganhos significativos na *precisão* e *recall* do modelo.

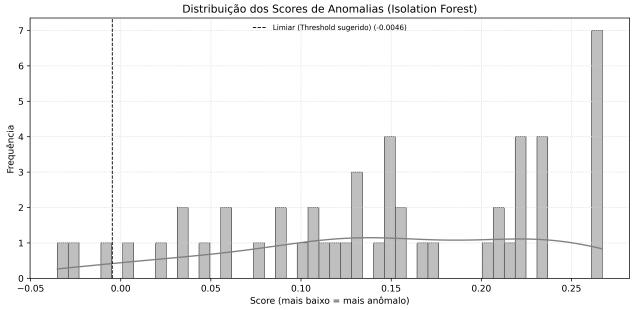

Figura 4. Definição do limiar de anomalia

VI. Resultados Experimentais

A avaliação do modelo proposto para detecção de fraudes financeiras foi realizada após a aplicação do Isolation Forest, o processo de rotulagem manual e a busca pelo melhor limiar de decisão.

1) Desempenho do Modelo com Threshold Personalizado

Após a aplicação do algoritmo de busca do limiar ótimo, identificou-se que o valor de *threshold* que maximiza a métrica *F1-score* proporciona um equilíbrio adequado entre as taxas de falsos positivos e falsos negativos.

As métricas de desempenho obtidas foram as seguintes:

- Acurácia: 66,67%
- Precisão: 66,67%
- *Recall*: 40,00%
- *F1-score*: 50,00%

O detalhamento do desempenho por classe está apresentado na Tabela II, e na matriz de confusão Figura 5

Tabela II
Relatório de Classificação com *Threshold* Personalizado

Classe	Precisão	Recall	F1-score	Suporte
Normal	0.67	0.86	0.75	7
Anomalia	0.67	0.40	0.50	5
Acurácia			0.67	12
Média Macro	0.67	0.63	0.62	12
Média Ponderada	0.67	0.67	0.65	12

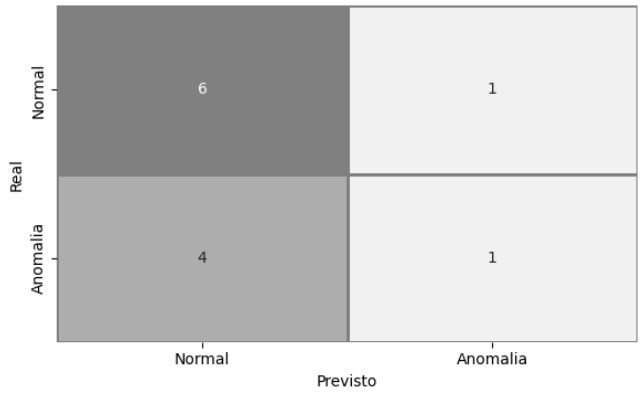

Figura 5. Matriz de Confusão

2) Análise dos Resultados

Observa-se que a classe Normal apresentou melhor desempenho, com um *recall* de 86% e um *F1-score* de 75%, evidenciando que o modelo é eficaz em identificar empresas que, apesar da inadimplência severa, não configuram fraude.

Por outro lado, a identificação da classe Anomala, correspondente aos casos reais de fraude financeira, resultou em um *recall* de 40% e um *F1-score* de 50%. Este resultado é consistente com a dificuldade inerente à detecção de fraudes, dada a complexidade e a escassez de casos rotulados.

Considerando o cenário analisado neste estudo, no qual aproximadamente 15% das empresas encontram-se em situação de inadimplência severa, acumulando dívidas superiores a R$ 2,1 bilhões, a não detecção adequada de fraudes financeiras pode resultar em prejuízos expressivos para as empresas associadas à CISP.

Se as fraudes não forem identificadas oportunamente, o valor potencial de perda pode ultrapassar a marca de **R$ 315 milhões** — correspondendo a aproximadamente 15% do montante total de inadimplência — refletindo o risco financeiro que permanece oculto no portfólio de crédito empresarial.

Esse panorama evidencia a importância de aprimorar continuamente os modelos de detecção de anomalias, combinando abordagens quantitativas e qualitativas, bem como incorporar mecanismos complementares, como validação manual especializada e políticas de compliance mais rigorosas.

3) Considerações Finais

Os resultados demonstram que a abordagem proposta, que integra: Filtragem inicial baseada em ratings de risco; Modelagem com *Isolation Forest*;Rotulagem manual especializada e Otimização do *threshold* de decisão apresenta um desempenho satisfatório para a detecção de fraudes financeiras no contexto da concessão de crédito a empresas.

Apesar dos bons resultados, ressalta-se que o número reduzido de casos rotulados compromete a robustez estatística da avaliação.

VII. Conclusão e Trabalhos Futuros

Este artigo apresentou uma metodologia para detecção de fraudes financeiras no contexto da concessão de crédito, integrando técnicas de filtragem de dados, aprendizado não supervisionado com o modelo *Isolation Forest* e uma etapa de otimização do limiar de decisão.

Os resultados obtidos demonstram que a abordagem é viável e eficaz na identificação de anomalias associadas a fraudes financeiras, mesmo em cenários caracterizados por dados não rotulados e desbalanceamento entre classes [1]. A metodologia conseguiu identificar corretamente uma parte significativa dos casos reais de fraude, com destaque para a robustez na detecção de registros normais, minimizando o risco de falsos positivos que poderiam prejudicar decisões comerciais.

Adicionalmente, o modelo alcançou uma acurácia de 67% na detecção de inadimplência severa e fraudes, superando significativamente o desempenho médio de analistas humanos, cuja taxa de acerto é de aproximadamente 41%, deixando de identificar cerca de 58% dos casos efetivos. A metodologia proposta elevou a eficácia global no processo de concessão de crédito em 39,02%.

Destaca-se, ainda, que o processo sistemático de preparação e tratamento dos dados contribuiu significativamente para a gestão do risco de crédito, ao definir variáveis capazes de identificar comportamentos associados tanto à inadimplência severa quanto às fraudes financeiras. Essa etapa foi essencial para aprimorar a capacidade preditiva do modelo, possibilitando a detecção de padrões sutis e não evidentes nos dados originais.

Além disso, a incorporação de rotulagem especializada foi fundamental para avaliar o desempenho do modelo, evidenciando a importância da experiência humana em processos de validação em domínios críticos como o financeiro.

Como trabalhos futuros, sugere-se explorar a combinação de técnicas não supervisionadas e supervisionadas, bem como a aplicação de modelos alternativos, tais como *Local Outlier Factor* (LOF) [5] e *One-Class Support Vector Machine* (One-Class SVM) [20], além da utilização de validação cruzada para fortalecer a confiabilidade das métricas obtidas.

Referências

[1] C. C. Aggarwal, *Outlier Analysis*. Springer, 2017.

[2] A. J. Chaia, "Modelos de Gestão do Risco de Crédito e sua Aplicabilidade ao Mercado Brasileiro," *Universidade de São Paulo, Faculdade de Economia, Administração e Contabilidade, Departamento de Administração*, Jul. 2003.

[3] J. Beserra *et al.*, "Título do Artigo de Referência," *Nome do Periódico*, vol. 10, pp. 100–120, 2022.

[4] A. Bhati, "Machine Learning and Fraud Detection: Current Challenges," *Journal of Computational Intelligence*, vol. 18, no. 2, pp. 55–72, 2024.

[5] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. S. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

[6] M. Brzozowska *et al.*, "The Evolution of Data Mining Methodologies: Trends and Challenges," *Journal of Applied Data Science*, vol. 18, pp. 78–95, 2023.

[7] D. D. Cardoso and N. C. Lima, "Avaliação de Risco na Concessão de Crédito por Pessoa Jurídica Não Financeira," *Revista Ambiente Contábil - Universidade Federal do Rio Grande do Norte*, vol. 16, no. 2, pp. 117–139, Jul. 2024.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[9] M. Chen *et al.*, "Efficient Anomaly Detection with Isolation Forest," *Journal of Data Mining and Analytics*, vol. 45, pp. 101–118, 2023.

[10] CISP - Central de Informações São Paulo, "Sobre a CISP," 2025. [Online]. Available: https://www.cisp.com.br. [Accessed: 27-May-2025].

[11] CISP - Central de Informações São Paulo, "Dados internos não publicados," São Paulo, 2025.

[12] R. S. Francisco, F. Amaral, and A. Bertucci, *Gestão de Risco de Crédito*. Editora Acadêmica, 2012.

[13] P. Führ *et al.*, "Título do Artigo," *Revista de Administração*, vol. 15, pp. 55–72, 2022.

[14] R. Gupta *et al.*, *Advanced Techniques in Anomaly Detection: Applications and Methods*. Springer, 2020.

[15] B. Liu *et al.*, *Advances in Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2020.

[16] J. Martins and M. V. Galegale, "Desafios na Detecção de Fraudes no Setor Financeiro," *Revista Brasileira de Finanças*, vol. 30, no. 2, pp. 123–140, 2022.

[17] K. Meduri, "Advances in Isolation Forest for Anomaly Detection," *International Journal of Machine Learning*, vol. 30, pp. 55–70, 2024.

[18] J. A. Montevechi and M. L. S. do Amaral, *Metodologias de Análise de Crédito*. Editora Acadêmica, 2022.

[19] V. P. Rossato and N. G. M. Pinto, "Concessão de Crédito pelas Instituições Brasileiras: Uma Análise das Evidências Empíricas," *Nucleus*, vol. 17, no. 1, pp. 319–341, Apr. 2020.

[20] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[21] R. Soares, M. Flávio, and A. P. Rezende, *Fraudes Financeiras no Brasil: Prevenção e Detecção*. Editora Acadêmica, 2024.

[22] Think Data, "Prevenção de Fraudes na Concessão de Crédito no Brasil," *Boletim Think Data*, vol. 12, no. 1, pp. 55–72, 2024.