UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de fraudes em seguros automotivos com aprendizado de máquina e inteligência artificial explicável (XAI)

Eduardo Barbante Rodrigues

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Eduardo Barbante Rodrigues

Detecção de fraudes em seguros automotivos com aprendizado de máquina e inteligência artificial explicável (XAI)

Monografia apresentada 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

Orientadora: Profa. Dra. Cibele Maria Russo Novelli

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Rodrigues, Eduardo Barbante
Detecção de fraudes em seguros automotivos com aprendizado de máquina e inteligência artificial explicável (XAI) / Eduardo Barbante Rodrigues ; orientadora Cibele Maria Russo Novelli. – São Carlos, 2025.
85 p. : il. (algumas color.) ; 30 cm.

Monografia - MBA em Inteligência Artificial e Big Data – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Novelli, Cibele Maria Russo, orient. II. Título.

Eduardo Barbante Rodrigues

Fraud detection in auto insurance with machine learning and explainable artificial intelligence (XAI)

Dissertation presented to the Departamento de Ciências de Computação of the Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Cibele Maria Russo Novelli

Original version

São Carlos

2025

Dedico este trabalho à minha esposa Letícia, pelo apoio constante, paciência e incentivo em todos os momentos. Ao meu filho Benício, que me inspira diariamente a ser uma versão melhor de mim mesmo. Aos colegas do MBA, pela troca de experiências, companheirismo e aprendizado compartilhado ao longo desta jornada. E à orientadora, pelo tempo e orientação dedicados à construção deste trabalho.

AGRADECIMENTOS

Agradeço à minha esposa Letícia, pelo amor, paciência e apoio incondicional durante toda esta jornada. Ao meu filho Benício, pela inspiração diária e por dar novo significado a cada conquista. Aos colegas do MBA em Inteligência Artificial e Big Data, pela parceria, pelas trocas e pelos aprendizados compartilhados que tornaram o caminho mais leve e enriquecedor.

À professora Solange Oliveira Rezende, pela empatia e incentivo em um momento de grandes desafios pessoais e profissionais — seu apoio foi essencial para que eu retomasse o foco e concluísse este trabalho com determinação.

À orientadora, pelas contribuições e pela orientação dedicada na condução deste estudo. E ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, pela excelência do programa e por proporcionar um ambiente de aprendizado tão transformador.

"A inteligência é a capacidade de se adaptar à mudança."
Stephen Hawking

RESUMO

RODRIGUES, E. B. **Detecção de fraudes em seguros automotivos com aprendizado de máquina e inteligência artificial explicável (XAI)**. 2025. 85 p. Monografia - MBA em Inteligência Artificial e Big Data - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

O setor de seguros é estratégico para a economia brasileira, mas enfrenta desafios relevantes devido à incidência de fraudes. Em 2024, o Sistema de Quantificação de Fraudes da CNseg registrou perdas de R$ 700,64 milhões em fraudes comprovadas no ramo de Automóveis, evidenciando a necessidade de aperfeiçoar os mecanismos de detecção. Este trabalho propõe e avalia um sistema de detecção de fraudes em seguros automotivos baseado em aprendizado de máquina, integrado a três perspectivas complementares: desempenho preditivo, viabilidade econômica e interpretabilidade das decisões.

Utiliza-se o conjunto de dados público *Fraud Oracle Dataset* (15,420 registros, 33 variáveis e cerca de 6% de fraudes). A metodologia envolve engenharia de atributos avançada — incluindo detecção de anomalias via *Isolation Forest* e *target encoding* — e técnicas de reamostragem baseadas em SMOTE para tratamento do desbalanceamento. São avaliados algoritmos de *Gradient Boosting* e *Random Forest*, com otimização bayesiana de hiperparâmetros via Optuna e ajuste de limiar de decisão. A seleção do modelo utiliza um *score* composto por MCC, G-Mean e Kappa, priorizando robustez em cenários desbalanceados.

Os resultados indicam a superioridade dos modelos de *gradient boosting* em relação aos *baselines* avaliados. O modelo selecionado (`CatBoost + SMOTEENN`) apresenta bom equilíbrio entre detecção de fraudes e controle de falsos positivos, além de viabilidade econômica robusta, com retorno sobre investimento significativamente superior às estratégias de referência. A análise de interpretabilidade via SHAP mostra que a variável de culpa do segurado é o preditor dominante, em consonância com a lógica de negócio do setor.

Conclui-se que a combinação de algoritmos de *ensemble* com técnicas de Inteligência Artificial Explicável (XAI) configura uma solução eficaz, economicamente viável e auditável, capaz de reduzir perdas financeiras e atender aos requisitos de governança e tomada de decisão das seguradoras.

Palavras-chave: Aprendizado de máquina. Detecção de fraudes. Seguros de automóveis. Dados desbalanceados. Inteligência artificial explicável (XAI). SHAP.

ABSTRACT

RODRIGUES, E. B. **Fraud detection in auto insurance with machine learning and explainable artificial intelligence (XAI)**. 2025. 85 p. Dissertation – MBA in Artificial Intelligence and Big Data - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

The insurance sector is strategic for the Brazilian economy but faces relevant challenges due to the incidence of fraud. In 2024, the Fraud Quantification System of CNseg recorded losses of R$ 700.64 million in confirmed fraud cases in the Auto line, highlighting the need to improve detection mechanisms. This work proposes and evaluates an automotive insurance fraud detection system based on machine learning, integrated into three complementary perspectives: predictive performance, economic viability, and interpretability of decisions.

We use the public *Fraud Oracle Dataset* (15,420 records, 33 variables, and about 6% fraud). The methodology involves advanced feature engineering — including anomaly detection via *Isolation Forest* and *target encoding* — and SMOTE-based resampling techniques to handle class imbalance. We evaluate *Gradient Boosting* and *Random Forest* algorithms, with Bayesian hyperparameter optimization via Optuna and decision-threshold tuning. Model selection relies on a composite score combining MCC, G-Mean, and Kappa, prioritizing robustness in imbalanced scenarios.

The results indicate that *gradient boosting* models outperform the evaluated baselines. The selected model (`CatBoost + SMOTEENN`) achieves a good balance between fraud detection and control of false positives, as well as robust economic viability, with return on investment significantly higher than reference strategies. The SHAP-based interpretability analysis shows that the policyholder fault variable is the dominant predictor, in line with the business logic of the sector.

We conclude that combining *ensemble* algorithms with Explainable Artificial Intelligence (XAI) techniques provides an effective, economically viable, and auditable solution capable of reducing financial losses and supporting governance and decision-making in insurance companies.

Keywords: Machine learning. Fraud detection. Auto insurance. Imbalanced data. Explainable artificial intelligence (XAI). SHAP.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABI	*Association of British Insurers*
ABNT	Associação Brasileira de Normas Técnicas
AUC	*Area Under the ROC Curve* (Área sob a Curva ROC)
CAIF	*Coalition Against Insurance Fraud*
CNseg	Confederação Nacional das Seguradoras
CV	*Cross-Validation* (Validação Cruzada)
EDA	*Exploratory Data Analysis* (Análise Exploratória de Dados)
FN	*False Negative* (Falso Negativo)
FP	*False Positive* (Falso Positivo)
G-Mean	*Geometric Mean* (Média Geométrica)
IA	Inteligência Artificial
ICMC	Instituto de Ciências Matemáticas e de Computação
Kappa	Coeficiente Kappa de Cohen
KNN	*K-Nearest Neighbors* (K-Vizinhos Mais Próximos)
LDA	*Linear Discriminant Analysis* (Análise Discriminante Linear)
MBA	*Master of Business Administration*
MCC	*Matthews Correlation Coefficient* (Coeficiente de Correlação de Matthews)
ML	*Machine Learning* (Aprendizado de Máquina)
MLP	*Multi-Layer Perceptron*
QDA	*Quadratic Discriminant Analysis* (Análise Discriminante Quadrática)
RF	*Random Forest*
ROC	*Receiver Operating Characteristic* (Característica de Operação do Receptor)
SHAP	*SHapley Additive exPlanations*

SMOTE	*Synthetic Minority Over-sampling Technique* (Técnica de Sobreamostragem Sintética da Minoria)
SQF	Sistema de Quantificação de Fraudes
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
TCC	Trabalho de Conclusão de Curso
TN	*True Negative* (Verdadeiro Negativo)
TP	*True Positive* (Verdadeiro Positivo)
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos
XAI	*eXplainable Artificial Intelligence* (Inteligência Artificial Explicável)
XGBoost	*eXtreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

O setor de seguros possui relevância estratégica para a economia brasileira, atuando na proteção financeira de indivíduos e empresas e mitigando riscos associados a diferentes atividades. Dentro desse contexto, o ramo de seguros de automóveis se destaca pelo elevado volume de apólices e pela frequência de sinistros, além de apresentar índices significativos de fraude, que impactam custos, preços e a eficiência operacional das seguradoras.

De acordo com o 22º Ciclo do Sistema de Quantificação de Fraudes (SQF) da CNseg, que apresenta os resultados consolidados de 2024, o mercado de seguros gerais registrou R$ 41,0 bilhões em sinistros ocorridos. Desse total, R$ 5,44 bilhões (13,3%) foram classificados como suspeitos e R$ 1,10 bilhão (2,7%) foram confirmados como fraude comprovada, indicando aumento do indicador de fraude em relação a 2023 e evidenciando a necessidade de aprimoramento dos mecanismos de detecção (CNseg, 2024).

No ramo Automóvel, os resultados são ainda mais expressivos. Em 2024, foram registrados R$ 29,30 bilhões em sinistros ocorridos, com R$ 4,02 bilhões (13,9%) classificados como suspeitos. As fraudes detectadas somaram R$ 910,66 milhões (3,2%), enquanto as fraudes comprovadas totalizaram R$ 700,64 milhões (2,4%). Em número de ocorrências, 30,5% dos sinistros automotivos foram considerados suspeitos e 23,5% resultaram em comprovação de fraude, demonstrando que o seguro de automóveis apresenta incidência significativamente elevada quando comparado aos demais ramos do mercado segurador brasileiro (CNseg, 2024).

Em mercados internacionais, observa-se comportamento semelhante. No Reino Unido, a *Association of British Insurers* (ABI) registrou, em 2022, 72.600 sinistros fraudulentos, totalizando £ 1,1 bilhão em perdas, com estimativa de que valor equivalente não seja detectado pelos mecanismos de monitoramento. O relatório indica que o seguro automotivo permanece como o ramo com maior frequência de fraudes naquele país (ABI, 2022).

Nos Estados Unidos, o estudo *The Impact of Insurance Fraud on the U.S. Economy*, desenvolvido pela *Coalition Against Insurance Fraud* (CAIF), estima perdas anuais de US$ 308,6 bilhões decorrentes de fraudes no setor. Dentro desse montante, o segmento *Property & Casualty* — que inclui o seguro automotivo — é responsável por aproximadamente US$ 45 bilhões, reforçando a magnitude do problema em escala internacional (CAIF, 2022).

Fraudes automotivas abrangem desde superestimação de danos e ocultação de informações até colisões simuladas e esquemas organizados. A ampliação da digitalização e a maior sofisticação dessas práticas tornam os métodos tradicionais de auditoria insuficientes,

exigindo a adoção de métodos analíticos avançados. Para investigar esse problema em um ambiente controlado e reprodutível, este estudo utiliza o conjunto de dados público *Vehicle Insurance Claim Fraud Detection* (também conhecido como *Fraud Oracle*), que reflete as características de desbalanceamento e complexidade observadas no mercado real.

1.2 Justificativa e motivação

A elevada incidência de fraudes no ramo de seguros de automóveis, evidenciada pelos indicadores da CNseg, torna esse problema crítico para o setor. Com fraudes comprovadas somando mais de R$ 700 milhões apenas em 2024, há uma necessidade financeira e operacional clara de implementar modelos preditivos mais assertivos, capazes de identificar padrões complexos em grandes volumes de dados (CNseg, 2024).

Nesse contexto, a justificativa para este estudo transcende a busca por precisão estatística e adentra a dimensão da **eficiência econômica**. No mercado de seguros, existe uma assimetria de custos relevante: o prejuízo financeiro de uma fraude não detectada (que pode corresponder ao valor integral de um veículo) é substancialmente superior ao custo operacional de uma sindicância ou investigação especializada. Portanto, a modelagem proposta justifica-se não apenas pela capacidade de classificar corretamente, mas pela necessidade de otimizar a alocação de recursos investigativos, buscando que o custo adicional de detecção seja compensado pela economia em sinistros indevidos. Essa perspectiva é posteriormente quantificada por meio de métricas de negócio, como Benefício Líquido (*Net Benefit*) e Retorno sobre o Investimento (ROI), discutidas nos capítulos metodológico e de resultados.

Sob a perspectiva prática, as seguradoras enfrentam o desafio de escolher entre múltiplas técnicas de modelagem para atingir essa eficiência. A avaliação comparativa entre modelos supervisionados — desde abordagens lineares, como a regressão logística, até *ensembles* de árvores mais complexos, como *Random Forest* e métodos de *Gradient Boosting* — é fundamental para identificar quais técnicas oferecem maior robustez diante de bases altamente desbalanceadas. Essa análise fornece evidências objetivas para apoiar decisões técnicas sobre qual arquitetura implementar em produção.

Do ponto de vista teórico e metodológico, este trabalho contribui ao integrar essa comparação de desempenho com técnicas de interpretabilidade, especificamente o SHAP (*SHapley Additive exPlanations*). A literatura enfatiza que, em domínios regulados, a "caixa-preta" dos modelos avançados é uma barreira para adoção. A incorporação de técnicas de Inteligência Artificial Explicável (XAI) permite compreender os fatores de risco, favorecendo a aderência às exigências de auditoria, governança e *compliance* do setor segurador.

Por fim, a utilização de um *dataset* público assegura a reprodutibilidade científica,

permitindo que outros pesquisadores validem os resultados, algo frequentemente inviável em estudos que utilizam dados proprietários protegidos por sigilo corporativo.

1.3 Problema de pesquisa e objetivos

A detecção de fraudes em seguros de automóveis constitui um desafio duplo. Primeiro, há o impacto financeiro direto e a dificuldade de identificação precisa, o que pode levar à negativa indevida de sinistros legítimos (falsos positivos) ou à aceitação de fraudes (falsos negativos). Segundo, existe o desafio metodológico do forte desbalanceamento entre as classes: tanto em bases reais quanto no conjunto *Vehicle Insurance Claim Fraud Detection*, os casos de fraude representam uma pequena fração dos registros, o que enviesa modelos tradicionais e torna métricas como a acurácia enganosas.

Embora técnicas de aprendizado de máquina, como *Gradient Boosting*, tenham se mostrado promissoras, muitos estudos concentram-se na maximização de métricas de performance sem endereçar adequadamente a interpretabilidade. Decisões de investigação de sinistro demandam argumentos compreensíveis. Modelos de alta complexidade tendem a ser opacos, criando uma lacuna entre o desempenho estatístico e a aplicabilidade no negócio.

Diante desse contexto, este trabalho delimita-se ao estudo experimental de classificação binária de fraudes, com foco na avaliação comparativa de modelos supervisionados sob três pilares principais: (i) tratamento do desbalanceamento por meio de técnicas de reamostragem; (ii) uso de métricas robustas de avaliação em bases desbalanceadas, combinadas ao ajuste de limiar de decisão e à quantificação de métricas de negócio; e (iii) interpretabilidade via SHAP aplicada ao modelo campeão. Busca-se responder não apenas qual modelo apresenta melhor desempenho estatístico e impacto financeiro potencial, mas também como seus resultados podem ser explicados de forma transparente para apoiar a tomada de decisão.

1.3.1 Objetivo geral

Avaliar, comparar e interpretar modelos de aprendizado de máquina supervisionado aplicados à detecção de fraudes em seguros de automóveis, considerando o forte desbalanceamento da base de dados e a necessidade de técnicas de explicabilidade.

1.3.2 Objetivos específicos

- Realizar a análise exploratória dos dados (*Exploratory Data Analysis* – EDA) para compreender a distribuição das variáveis, identificar padrões e quantificar o desbalanceamento severo da base.

- Desenvolver um pipeline de engenharia de atributos (*feature engineering*) robusto, incorporando técnicas de detecção de anomalias (*Isolation Forest*) e criação de variáveis de risco e interação, visando enriquecer o poder preditivo dos dados.
- Configurar um pipeline experimental customizado em Python, integrando *Scikit-Learn* e *Imbalanced-Learn*, empregando técnicas avançadas de reamostragem (incluindo SMOTE, ADASYN e métodos híbridos) para mitigar o desbalanceamento, assegurando o controle rigoroso de vazamento de dados.
- Implementar e otimizar algoritmos baseados em *Gradient Boosting* (XGBoost, LightGBM e CatBoost) e *Random Forest*, utilizando otimização Bayesiana de hiperparâmetros (via Optuna) e ajuste de limiar de decisão (*threshold tuning*) focados na maximização do Kappa de Cohen.
- Avaliar o desempenho dos modelos priorizando métricas robustas ao desbalanceamento, com ênfase no *Matthews Correlation Coefficient* (MCC), na *Geometric Mean* (G-mean) e no Kappa de Cohen.
- Quantificar o impacto financeiro dos modelos propostos por meio de métricas de negócio, estimando o Benefício Líquido e o Retorno sobre o Investimento (ROI) das investigações em um cenário simulado.
- Aplicar técnicas de Interpretabilidade (*Explainable AI*), utilizando SHAP (*SHapley Additive exPlanations*), para identificar os principais fatores de risco e garantir a transparência das decisões do modelo.
- Discutir criticamente os resultados e a interpretabilidade dos modelos, destacando as implicações financeiras, as vantagens operacionais e as limitações da abordagem proposta frente aos desafios reais das seguradoras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo estabelece a fundamentação teórica que sustenta esta pesquisa. A revisão da literatura está estruturada em um formato de funil, iniciando com o domínio do problema — a fraude em seguros (Seção 2.1) — e seguindo para as soluções computacionais, abordando o aprendizado de máquina supervisionado (Seção 2.2). Em seguida, aprofunda-se nos desafios técnicos centrais do trabalho: o tratamento de dados desbalanceados (Seção 2.3) e a seleção de métricas de avaliação adequadas (Seção 2.4). Por fim, discute-se a interpretabilidade de modelos com foco em SHAP (Seção 2.5) e analisam-se trabalhos correlatos (Seção 2.6), posicionando esta pesquisa no contexto científico atual.

2.1 Fraude em seguros de automóveis

A fraude em seguros pode ser entendida como a prática deliberada de enganar, ocultar ou distorcer informações relevantes com o objetivo de obter um benefício financeiro indevido de uma apólice de seguro. Essa definição é coerente com a concepção apresentada por (Derrig, 2002), que caracteriza fraude como atos intencionais e potencialmente criminosos destinados a gerar pagamentos não devidos. Este fenômeno não se restringe a um único ramo, mas, no setor de seguros de automóveis, assume contornos particularmente significativos, conforme evidenciado pelos dados apresentados no Capítulo 1.

Relatórios de entidades como a Confederação Nacional das Seguradoras (CNseg), a *Association of British Insurers* (ABI) e a *Coalition Against Insurance Fraud* (CAIF) indicam que o seguro automotivo concentra consistentemente uma parcela elevada das perdas financeiras e da frequência de ocorrências fraudulentas (CNseg, 2024; ABI, 2022; CAIF, 2022).

A literatura e os relatórios do setor classificam essas fraudes em diversas tipologias, que, como observado na Introdução (Capítulo 1), incluem:

- **Fraudes de oportunidade (ou *padding*):** superestimação de danos reais após um sinistro legítimo, incluindo partes do veículo não afetadas pelo acidente.
- **Sinistros simulados (ou *staged accidents*):** criação de colisões deliberadas, muitas vezes envolvendo múltiplos veículos e conluio entre os participantes, para registrar um sinistro falso.
- **Fraudes documentais e de subscrição (ou *premium fraud*):** omissão ou falsificação de informações no momento da contratação da apólice (por exemplo, endereço ou perfil do condutor principal) para obter um prêmio (*preço*) reduzido.

- **Fraudes de esquemas organizados:** operações complexas envolvendo redes de indivíduos, que podem incluir oficinas, corretores e até mesmo profissionais de saúde, com o objetivo de fraudar seguradoras em larga escala.

O impacto financeiro bilionário, demonstrado pelos dados da CNseg e da CAIF (Capítulo 1), pressiona as seguradoras a buscar mecanismos de detecção mais eficientes, pois esses custos são, em última análise, repassados aos prêmios de todos os segurados.

2.2 Aprendizado de máquina supervisionado na detecção de fraude

Historicamente, a detecção de fraude dependia de revisões manuais e sistemas baseados em regras de negócio. Esses sistemas, embora úteis, são estáticos, fáceis de burlar por fraudadores que aprendem os padrões e incapazes de identificar esquemas complexos ou emergentes (Bolton; Hand, 2002).

A evolução para métodos estatísticos, como a regressão logística, permitiu uma abordagem mais quantitativa, mas foi com o advento do aprendizado de máquina que o campo ganhou novas capacidades. O aprendizado de máquina supervisionado, em particular, tornou-se a abordagem dominante para problemas de classificação de fraude (Abdallah; Maarof; Zainal, 2016).

Nessa abordagem, o algoritmo é treinado com um conjunto de dados históricos em que cada sinistro (ocorrência) já foi rotulado como "fraude" (classe minoritária) ou "não fraude" (classe majoritária). O objetivo do modelo é aprender padrões complexos e interações entre as variáveis (por exemplo, hora do sinistro, tipo de veículo, perfil do condutor, histórico de sinistros) que distingam um caso fraudulento de um legítimo.

Como apontado na Introdução (Capítulo 1), diversos algoritmos são explorados na literatura para essa tarefa:

- **Modelos lineares (por exemplo, regressão logística):** usados como *baseline* por sua interpretabilidade, mas geralmente com menor poder preditivo em problemas complexos.

- **Modelos baseados em árvores (por exemplo, árvores de decisão, *Random Forest*):** populares pela capacidade de capturar relações não lineares e pela robustez. O *Random Forest*, um método de *ensemble*, mitiga o superajuste (*overfitting*) das árvores de decisão individuais.

- **Modelos de *gradient boosting* (por exemplo, XGBoost, LightGBM):** frequentemente apresentam o estado da arte em desempenho para dados tabulares, como os de seguros. Eles constroem modelos de forma sequencial, em que cada nova árvore corrige os erros da anterior, resultando em alta capacidade preditiva.

O foco da aplicação de aprendizado de máquina supervisionado na detecção de fraude não é, portanto, criar um novo algoritmo, mas avaliar e comparar sistematicamente as abordagens existentes para determinar quais oferecem melhor capacidade de generalização e discriminação no contexto específico dos sinistros de automóveis.

Contudo, a capacidade de generalização desses algoritmos, especialmente os métodos de *ensemble* e *boosting*, é fortemente dependente da configuração de seus hiperparâmetros. A utilização de configurações padrão (*default*) frequentemente subutiliza o potencial preditivo dos modelos, sendo insuficiente para cenários complexos como a detecção de fraudes. Portanto, a comparação justa entre algoritmos exige que cada um seja otimizado para extrair seu máximo desempenho, o que demanda técnicas de busca mais sofisticadas que a tentativa e erro manual.

2.2.1 Otimização de Hiperparâmetros

Diferente dos parâmetros internos aprendidos pelo modelo durante o treinamento (como os pesos em uma regressão), os hiperparâmetros definem a estrutura e o comportamento do algoritmo a priori (por exemplo, a profundidade máxima de uma árvore ou a taxa de aprendizado do *gradient boosting*).

Métodos tradicionais de busca por hiperparâmetros, como *Grid Search* (busca em grade) e *Random Search* (busca aleatória), embora populares, muitas vezes são computacionalmente ineficientes por não utilizarem o conhecimento das tentativas anteriores. Abordagens mais robustas utilizam Otimização Bayesiana, como o algoritmo *Tree-structured Parzen Estimator* (TPE), implementado na biblioteca Optuna (Akiba *et al.*, 2019). O TPE modela a probabilidade de um conjunto de hiperparâmetros gerar um bom resultado baseando-se no histórico de avaliações, concentrando a busca nas regiões mais promissoras do espaço de parâmetros. Isso permite encontrar configurações ótimas com menor custo computacional, sendo essencial para o ajuste fino de modelos complexos como o XGBoost.

2.3 Desbalanceamento de dados em problemas de fraude

O desafio central na aplicação de aprendizado de máquina à detecção de fraude, como enfatizado no problema de pesquisa (Capítulo 1), é o desbalanceamento de classes. Por definição, a fraude é um evento raro. Em um portfólio de seguros, a vasta maioria dos sinistros é legítima, e os casos fraudulentos representam apenas uma fração do total de registros.

Esse desbalanceamento tem consequências severas para o treinamento de modelos. Algoritmos de aprendizado de máquina são tipicamente otimizados para minimizar o erro global (como a acurácia). Em uma base com 99% de casos legítimos e 1% de fraude, um modelo que classifica todos os sinistros como legítimos alcançará 99% de acurácia, embora seja completamente inútil para o propósito de detecção (Altalhan; Algarni, 2025).

O modelo desenvolve um viés para a classe majoritária, falhando em aprender os padrões sutis da classe minoritária (fraude).

A literatura de *imbalanced learning* (aprendizado em dados desbalanceados) propõe diversas estratégias para mitigar esse problema, que podem ser agrupadas em três categorias principais (Krawczyk, 2016):

- **Nível de dados (amostragem):** modifica-se o conjunto de treinamento para criar uma distribuição de classes mais equilibrada.
  - **Undersampling (subamostragem):** remove aleatoriamente exemplos da classe majoritária (não fraude). Vantagem: reduz o custo computacional. Desvantagem: pode descartar informações úteis da classe majoritária.
  - **Oversampling (sobreamostragem):** duplica aleatoriamente exemplos da classe minoritária (fraude). Vantagem: não perde informação. Desvantagem: pode levar a *overfitting*, pois o modelo aprende padrões de exemplos idênticos.
  - **SMOTE (*Synthetic Minority Over-sampling Technique*):** conforme destacado nos objetivos deste TCC (Capítulo 1), o SMOTE é uma técnica avançada de *oversampling*. Proposto por (Chawla *et al.*, 2002), em vez de duplicar registros, ele cria novos exemplos sintéticos da classe minoritária, interpolando entre vizinhos próximos existentes. Isso permite ao modelo aprender regiões de decisão mais amplas e robustas para a classe de fraude.
  - **Variações do SMOTE (ADASYN, SMOTEENN, SMOTETomek):** além do SMOTE clássico, a literatura propõe extensões que refinam a forma de gerar ou limpar amostras. O ADASYN (*Adaptive Synthetic Sampling*) foca a geração de exemplos sintéticos nas regiões em que a classe minoritária é mais difícil de aprender, atribuindo mais peso a instâncias com maior dificuldade de classificação (He *et al.*, 2008). Já métodos híbridos como SMOTEENN e SMOTETomek combinam *oversampling* sintético com técnicas de limpeza da classe majoritária (remoção de exemplos ambíguos ou ruidosos), com o objetivo de reduzir sobreposição entre classes e melhorar a definição das fronteiras de decisão (Batista; Prati; Monard, 2004).
- **Nível de algoritmo:** modifica-se o algoritmo de aprendizado para dar mais importância à classe minoritária (por exemplo, ajuste de pesos de classe).
- **Abordagens híbridas:** combinam técnicas de amostragem com modificações algorítmicas.

A escolha de técnicas baseadas em SMOTE, centrais para este TCC, busca resolver o viés de treinamento na origem, fornecendo ao modelo uma visão mais equilibrada do

problema antes mesmo da otimização. No presente estudo, essas técnicas incluem tanto o SMOTE clássico quanto variações como ADASYN e métodos híbridos (SMOTEENN e SMOTETomek), exploradas experimentalmente em combinação com diferentes algoritmos de classificação.

2.4 Métricas de avaliação para dados desbalanceados

O segundo pilar do desafio do desbalanceamento é a avaliação. Como visto, a acurácia é uma métrica enganosa. A literatura aponta que mesmo a *AUC-ROC* (*Area Under the Receiver Operating Characteristic Curve*), embora mais robusta que a acurácia, pode ser excessivamente otimista em cenários de desbalanceamento extremo, pois considera a taxa de falsos positivos, que é naturalmente baixa quando a classe negativa é massiva (Huayanay; Bazán; Russo, 2024).

Para uma avaliação fidedigna da capacidade do modelo em detectar a classe minoritária (fraude) sem gerar um volume excessivo de falsos positivos (alarmes falsos, que custam caro para a operação de sinistros), métricas mais adequadas são necessárias. Este TCC, alinhado à literatura recente, foca em métricas que consideram explicitamente o equilíbrio entre as classes:

- **F1-score:** média harmônica entre precisão (dos casos classificados como fraude, quantos realmente são fraude) e *recall*/sensibilidade (de todas as fraudes reais, quantas o modelo capturou). É útil, mas pode ignorar o desempenho nos verdadeiros negativos.
- **G-mean (*Geometric Mean*):** raiz quadrada do produto da sensibilidade (*recall* da classe positiva) e da especificidade (*recall* da classe negativa). Uma G-mean baixa indica que o modelo está falhando em uma das classes (geralmente a minoritária).
- **Kappa de Cohen:** mede a concordância entre a previsão do modelo e a realidade, corrigindo a concordância que ocorreria apenas pelo acaso.
- **MCC (*Matthews Correlation Coefficient*):** conforme destacado nos objetivos (Capítulo 1), o MCC é considerado por muitos autores como uma das métricas mais robustas e completas para classificação binária desbalanceada (Chicco; Jurman, 2020). Ele varia de -1 (discordância total) a +1 (concordância perfeita), com 0 indicando desempenho aleatório. Sua força reside no fato de utilizar todos os quatro quadrantes da matriz de confusão (verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos), produzindo uma pontuação alta apenas se o modelo for bom em todas as classes.

O estudo de (Huayanay; Bazán; Russo, 2024), referência central para este trabalho, investiga o desempenho de diversas métricas em diferentes proporções de desbalanceamento.

Os autores demonstram empiricamente que métricas como a acurácia são insensíveis ao desempenho da classe minoritária, enquanto G-mean e MCC apresentam correlação mais forte com a capacidade do modelo de discriminar corretamente a classe de interesse, mesmo quando ela é extremamente rara. Além disso, o estudo atribui especial relevância ao Kappa de Cohen, ao lado da G-mean e do MCC, ao avaliar métricas em cenários de forte desbalanceamento. Em consonância com essa evidência, este TCC utiliza o MCC como métrica central de otimização e combina MCC, G-mean e Kappa em um *score* composto para avaliar e selecionar os modelos.

2.4.1 Métricas de negócio em detecção de fraude

Além das métricas estatísticas de classificação, problemas de detecção de fraude em seguros exigem a consideração explícita de métricas de negócio. Cada decisão de investigação gera custos (por exemplo, horas de analistas, vistorias presenciais, consultas a bancos de dados externos), ao mesmo tempo em que a identificação correta de fraudes evita perdas financeiras significativas associadas ao pagamento indevido de sinistros.

Nesse contexto, é comum definir indicadores que combinem a matriz de confusão com pressupostos de custo e benefício por caso. Entre as métricas de negócio relevantes para este trabalho, destacam-se:

- **Taxa de captura de fraude (*fraud catch rate*):** corresponde à proporção de fraudes efetivamente identificadas pelo modelo, isto é, a sensibilidade aplicada ao contexto de fraude. Em termos da matriz de confusão, é dada por $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, onde TP são fraudes corretamente classificadas e FN são fraudes não detectadas.
- **Benefício Líquido (*Net Benefit*):** estima o ganho financeiro resultante do uso do modelo, aproximando o valor total das fraudes evitadas menos o custo das investigações conduzidas. Em termos simplificados, pode ser expresso como

  $$\text{Net Benefit} \approx \mathrm{TP} \cdot C_{\text{fraude}} - (\mathrm{TP}+\mathrm{FP}) \cdot C_{\text{investigacao}},$$

  em que C_{fraude} representa o custo médio de uma fraude evitada e $C_{\text{investigacao}}$ o custo médio de investigar um sinistro sinalizado.
- **Retorno sobre o Investimento (ROI):** relaciona o benefício líquido ao custo total das investigações, fornecendo uma medida relativa de retorno financeiro:

  $$\mathrm{ROI} = \frac{\text{Net Benefit}}{(\mathrm{TP}+\mathrm{FP}) \cdot C_{\text{investigacao}}}.$$

A combinação de métricas técnicas robustas (como MCC, G-mean e Kappa) com métricas de negócio como Benefício Líquido e ROI permite avaliar não apenas a qualidade estatística do classificador, mas também seu impacto econômico potencial para a seguradora, alinhando o modelo preditivo aos objetivos práticos do negócio.

2.5 Interpretabilidade e explicabilidade de modelos (XAI) – foco em SHAP

Modelos de *boosting* como o XGBoost, embora potentes em predição, são frequentemente criticados como "caixas-pretas" (*black boxes*). Em um setor regulado como o de seguros, tomar decisões sensíveis (como negar um sinistro ou encaminhá-lo para investigação especial) com base em um modelo opaco é arriscado do ponto de vista de auditoria, governança e conformidade (Arrieta *et al.*, 2020).

Surge, assim, o campo da IA explicável (XAI, do inglês *Explainable AI*), que visa desenvolver técnicas para tornar as previsões de modelos complexos compreensíveis para humanos. Um *review* sistemático recente da literatura de XAI analisou aplicações da técnica e destacou que SHAP e LIME estão entre os métodos mais utilizados na prática. Esse *review* aponta que o SHAP é frequentemente preferido por suas garantias matemáticas mais robustas, o que justifica sua escolha para este TCC (Saarela; Podgorelec, 2024).

O SHAP (*Shapley Additive Explanations*), introduzido por (Lundberg; Lee, 2017), consolidou-se como um padrão para dados tabulares. Ele se baseia nos valores de Shapley, um conceito da teoria dos jogos cooperativos. A ideia central é tratar a previsão de um modelo como um "jogo" e as variáveis (*features*) como "jogadores". O SHAP calcula a contribuição marginal de cada "jogador" (variável) para o resultado final da "partida" (a predição de fraude), distribuindo de forma justa o "prêmio" (a diferença entre a predição do modelo e a predição média da base).

Em termos práticos, o SHAP permite responder a duas perguntas cruciais que este TCC busca explorar (Capítulo 1):

- **Interpretabilidade global:** quais variáveis (por exemplo, `valor_do_sinistro`, `idade_condutor`, `tempo_ate_notificacao`) são, em média, as mais importantes para o modelo tomar suas decisões de fraude?
- **Interpretabilidade local:** para um sinistro específico que foi classificado como fraude, quais foram os fatores que levaram o modelo a essa conclusão? Por exemplo, o `valor_do_sinistro` pode ter elevado a suspeita, enquanto a `idade_condutor` pode tê-la reduzido.

A aplicação do SHAP é, portanto, essencial para validar se o modelo está aprendendo padrões de fraude relevantes ou se está se baseando em *proxies* ou vieses indesejados nos dados, contribuindo para a transparência exigida pelo setor.

2.6 Trabalhos correlatos

A detecção de fraude em seguros com aprendizado de máquina é um campo de pesquisa ativo. Diversos trabalhos abordam peças do problema que este TCC se propõe a integrar.

Itri *et al.* realizaram um estudo comparativo de dez algoritmos de aprendizado de máquina para detecção de fraude em seguros de automóveis, utilizando o dataset público `carclaims.txt` (15.420 registros, 6% de fraudes). Os autores avaliaram modelos como J48, Naive Bayes, Random Forest e Multilayer Perceptron, empregando validação cruzada com 10 *folds* e balanceamento via reamostragem, reportando o Random Forest como melhor modelo segundo a K-Measure (99,46%). No entanto, a avaliação concentrou-se em F-Measure e K-Measure, sem explorar métricas mais robustas para desbalanceamento, como MCC, G-mean ou Kappa de Cohen. Este TCC se aproxima ao realizar uma comparação de modelos supervisionados em um *dataset* público de seguro automotivo com proporção semelhante de fraudes (Fraud Oracle), mas se diferencia pelo foco em métricas robustas para desbalanceamento (MCC, G-mean e Kappa) validadas pela literatura recente (Huayanay; Bazán; Russo, 2024) e pela camada de interpretação com SHAP, ausente no trabalho de Itri et al.

Ding *et al.* propuseram um *framework* PSO-XGBoost para detecção de fraude em seguros automotivos, combinando o algoritmo de otimização por enxame de partículas (PSO) com XGBoost para ajuste de hiperparâmetros. Os autores utilizaram dados privados de agências investigativas e instituições de seguros, além de dados do Kaggle, e reportaram acurácia de 95%, superior a métodos tradicionais (SVM, Naive Bayes, Logistic Regression) e ao Random Forest. Crucialmente, também aplicaram SHAP para interpretar as predições, identificando o valor do sinistro como principal *feature* de risco. Esse trabalho é próximo em escopo ao presente TCC, mas difere em três aspectos: (i) utiliza dados predominantemente privados, enquanto este TCC trabalha exclusivamente com o Fraud Oracle Dataset público, assegurando reprodutibilidade; (ii) adota a acurácia como métrica principal, ao passo que este TCC enfatiza métricas robustas ao desbalanceamento (MCC, G-mean e Kappa), com o MCC como métrica central de otimização e base de um *score* composto; (iii) não discute explicitamente o uso de técnicas de balanceamento, enquanto este TCC integra, em pipelines de modelagem, técnicas de reamostragem baseadas em SMOTE (SMOTE, ADASYN, SMOTEENN e SMOTETomek), aplicadas apenas sobre os dados de treino de cada partição, de modo a mitigar o desbalanceamento sem introduzir *data leakage*.

Komsrimorakot *et al.* propuseram uma técnica híbrida que integra Center Point SMOTE (CP-SMOTE) com Random Under-Sampling (RUS) para tratar o desbalanceamento de classes em sinistros de seguros automotivos. Os autores mostram que a combinação CP-SMOTE + RUS é mais eficaz que o SMOTE isolado para melhorar o equilíbrio entre classes. Este TCC se aproxima ao abordar explicitamente o desbalancea-

mento com técnicas baseadas em SMOTE, mas se diferencia por utilizar o SMOTE padrão e suas variações (ADASYN, SMOTEENN e SMOTETomek) integradas a pipelines com *ensembles* de árvores e por incorporar análise de explicabilidade com SHAP, ausente no trabalho de Komsrimorakot et al.

A lacuna que este TCC preenche, portanto, não reside na aplicação isolada de aprendizado de máquina ou do SHAP, mas na integração sistemática desses componentes. A contribuição principal consiste na avaliação rigorosa em um dataset público (Fraud Oracle), enfrentando o desbalanceamento com um pipeline tecnicamente robusto — que aplica variações de SMOTE (como ADASYN e SMOTETomek) exclusivamente no treino para evitar *data leakage*. Além disso, o estudo inova ao utilizar um score composto de métricas resilientes (MCC, G-mean, Kappa) como critério de decisão, validando o modelo final com uma camada de interpretabilidade via SHAP. Por fim, o trabalho documenta a resolução de incompatibilidades técnicas entre bibliotecas recentes, oferecendo uma solução prática e reprodutível para a comunidade.

3 METODOLOGIA

3.1 Tipo de Pesquisa

Esta pesquisa classifica-se metodologicamente segundo três dimensões: **natureza**, **objetivos** e **procedimentos técnicos**, conforme recomendações da literatura de metodologia científica.

3.1.1 Quanto à natureza

Esta pesquisa caracteriza-se como **Pesquisa Aplicada**, pois visa propor e avaliar soluções práticas para um problema real: a detecção de fraudes em sinistros de seguros de automóveis. As técnicas de aprendizado de máquina e interpretabilidade são empregadas com o objetivo de gerar conhecimento diretamente utilizável pelo setor segurador.

3.1.2 Quanto aos objetivos

Este trabalho configura-se como **Pesquisa Explicativa**. Além de avaliar o desempenho de diferentes algoritmos supervisionados, busca-se compreender *por que* determinados modelos apresentam melhor desempenho e *como* variáveis específicas influenciam a probabilidade de fraude, utilizando métodos de interpretabilidade como SHAP.

3.1.3 Quanto aos procedimentos técnicos

A pesquisa combina **Pesquisa Bibliográfica** e **Pesquisa Experimental**. A fundamentação teórica abrange temas como fraude em seguros, aprendizado supervisionado, desbalanceamento de classes, métricas de avaliação e interpretabilidade.

A etapa experimental envolve a avaliação controlada de algoritmos, técnicas de pré-processamento e métricas, seguindo boas práticas como a divisão estratificada dos dados em conjuntos de treino, validação e teste, bem como a utilização de validação cruzada estratificada na avaliação de robustez do modelo campeão.

3.2 Caracterização do Conjunto de Dados

O presente estudo utiliza o *Fraud Oracle Dataset*, uma base pública disponibilizada na plataforma Kaggle[1] e amplamente referenciada na literatura de detecção de fraudes em seguros automotivos. Trabalhos como o de Itri *et al.* (2019), por exemplo, utilizaram a versão clássica dessa base (`carclaims.txt`) para avaliar o desempenho de algoritmos

[1] Disponível em: https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection. Acesso em: 17 nov. 2025.

de aprendizado de máquina, o que reforça sua relevância acadêmica e sua adequação a estudos comparativos.

A escolha deste *dataset* fundamenta-se em quatro fatores essenciais. Primeiro, por ser uma base pública, garante a reprodutibilidade dos experimentos, permitindo que outros pesquisadores repliquem ou ampliem os resultados. Segundo, elimina restrições éticas e legais impostas pela LGPD, uma vez que não envolve dados sensíveis de segurados reais. Terceiro, contorna a indisponibilidade de bases proprietárias de seguradoras, protegidas por sigilo corporativo e regulatório. Por fim, trata-se de um conjunto supervisionado, contendo rótulos explícitos de fraude, o que o torna adequado à avaliação de modelos preditivos de classificação.

O conjunto de dados é estruturado em formato tabular e contém 15.420 registros descritos por 33 atributos. Cada registro representa um sinistro de seguro automotivo e inclui informações relacionadas ao veículo, ao segurado, ao histórico da apólice, às circunstâncias do acidente e aos detalhes da reclamação. A variável-alvo, denominada `FraudFound_P`, assume os valores 0 (sinistro legítimo) e 1 (sinistro fraudulento), caracterizando o problema como uma tarefa de classificação binária.

A distribuição das classes evidencia um desbalanceamento substancial. Dos 15.420 registros, 14.497 (94,01%) correspondem a sinistros legítimos, enquanto apenas 923 (5,99%) representam fraudes. Essa configuração resulta em uma razão aproximada de 15,7:1 (não fraude : fraude), refletindo a realidade operacional do mercado de seguros. Esse cenário justifica a adoção de técnicas avançadas de reamostragem — abrangendo o SMOTE e suas variações (como ADASYN e métodos híbridos) — bem como o uso de métricas mais robustas que a acurácia, tais como MCC, G-mean e Kappa, discutidas nas seções subsequentes.

Quanto à tipologia, o conjunto de dados é composto por 24 variáveis categóricas e 8 variáveis numéricas. Entre as categóricas, destacam-se `Make` (fabricante), `AccidentArea` (local do acidente) e `VehiclePrice` (faixa de preço do veículo). É importante notar que a variável `VehiclePrice` é apresentada em faixas discretas (por exemplo, "*more than 69000*") e não como um valor monetário contínuo. Além disso, o valor do veículo não coincide necessariamente com o valor da indenização paga em caso de sinistro, que pode ser parcial ou total, e o *dataset* não contém uma variável específica para o valor efetivo da indenização.

Essa combinação de fatores impõe uma limitação para o cálculo exato do prejuízo individual de cada sinistro, motivando a adoção, nas análises financeiras deste trabalho, de um custo médio fixo de fraude, conforme detalhado posteriormente na Seção 3.5.7. A lista completa dos atributos encontra-se no Apêndice .1.

3.3 Preparação dos Dados e Engenharia de Atributos

A preparação dos dados seguiu um pipeline estruturado e reprodutível, composto pelas etapas de limpeza do conjunto original, engenharia de atributos, transformação das variáveis e divisão estratificada em subconjuntos de treino, validação e teste. Todas essas operações foram implementadas manualmente em Python, utilizando as bibliotecas *pandas*, *scikit-learn* e *imbalanced-learn*, em conjunto com uma classe customizada de engenharia de atributos (*AdvancedFeatureEngineering*).

Na etapa de engenharia de atributos, o objeto *AdvancedFeatureEngineering* é ajustado exclusivamente sobre o conjunto de treino e, em seguida, aplicado aos conjuntos de validação e teste, preservando a separação entre as partições e evitando *data leakage*. As transformações subsequentes são organizadas por meio de um *ColumnTransformer*, que realiza a codificação *one-hot* das variáveis categóricas e repassa as variáveis numéricas sem normalização adicional. Por fim, técnicas de reamostragem baseadas em SMOTE (SMOTE, ADASYN, SMOTEENN e SMOTETomek) são encapsuladas, juntamente com os modelos de classificação, em *pipelines* do *scikit-learn*, de forma que a reamostragem ocorra apenas sobre os dados de treino em cada partição. A Figura 1 resume, de maneira esquemática, esse fluxo de preparação dos dados.

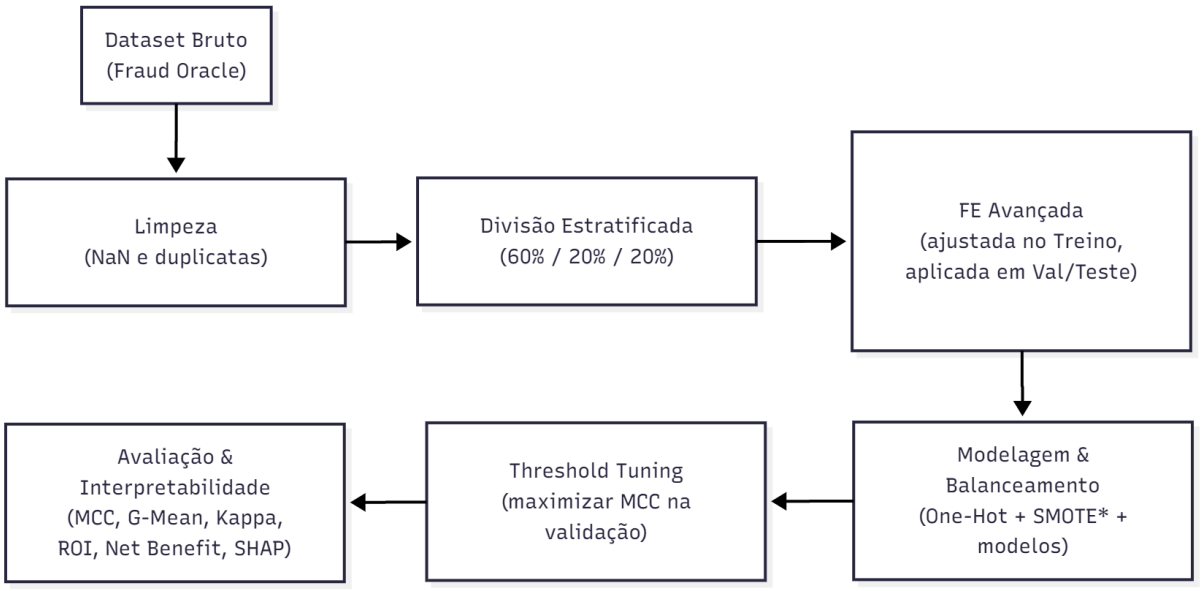

Figura 1 – Visão geral do pipeline de preparação dos dados e modelagem. A partir do *dataset* bruto (Fraud Oracle), realiza-se a limpeza dos dados, a divisão estratificada em treino, validação e teste (60%, 20% e 20%), a aplicação de engenharia de atributos avançada (incluindo *Isolation Forest* para geração de *anomaly score*, *risk score* e variáveis de interação), a modelagem com balanceamento de classes (codificação *one-hot* e técnicas de reamostragem baseadas em SMOTE* integradas a *pipelines*), o ajuste de limiar para maximizar o MCC na validação e, por fim, a avaliação no conjunto de teste com métricas robustas (MCC, G-Mean, Kappa), métricas de negócio (ROI e *Net Benefit*) e interpretabilidade via SHAP. SMOTE* refere-se ao conjunto de técnicas SMOTE, ADASYN, SMOTEENN e SMOTETomek. O conjunto de modelos avaliados inclui algoritmos otimizados via Optuna (XGBoost, LightGBM, CatBoost, Random Forest) e *baselines* (Regressão Logística).

Fonte: Elaborada pelo autor (2025).

3.3.1 Limpeza dos dados

A primeira etapa do pipeline consistiu na verificação da qualidade do *Fraud Oracle Dataset*, com foco na presença de valores ausentes e registros duplicados. A análise exploratória confirmou que as 33 variáveis originais não apresentam valores ausentes, o que dispensa procedimentos de imputação para este conjunto de dados. Da mesma forma, não foram identificados registros duplicados, uma vez que o atributo `PolicyNumber` atua como identificador único das apólices.

Além dessa verificação, foi realizada a remoção explícita da coluna `PolicyNumber` antes das etapas de engenharia de atributos e modelagem. Por se tratar de um identificador puramente administrativo, sem significado preditivo direto, o uso desse campo como *feature* poderia introduzir *data leakage* e levar o modelo a memorizar padrões artificiais associados

a códigos específicos de apólice. Assim, os 15.420 registros foram integralmente mantidos no experimento, porém com `PolicyNumber` excluído do conjunto de variáveis explicativas.

3.3.2 Transformação de variáveis

O dataset é composto por 24 atributos categóricos e 8 numéricos (excluindo a variável-alvo), o que exige transformações específicas para adequação aos algoritmos supervisionados e à construção dos *pipelines* de modelagem.

Codificação de variáveis categóricas. As variáveis categóricas são convertidas em representações numéricas por meio de *One-Hot Encoding*, implementado via *ColumnTransformer* do *scikit-learn*. Nesse arranjo, as colunas categóricas são encaminhadas para um *OneHotEncoder* configurado com `handle_unknown="ignore"`, enquanto as colunas numéricas são simplesmente repassadas (*passthrough*). O *ColumnTransformer* é ajustado apenas sobre o conjunto de treino e posteriormente aplicado às partições de validação e teste, garantindo que o espaço de features seja consistente entre os conjuntos sem introduzir *data leakage*. Essa codificação gera um conjunto expandido de variáveis binárias a partir das categorias originais, preservando a informação sem impor ordenação artificial.

Tratamento das variáveis numéricas. No pipeline final adotado neste trabalho, as variáveis numéricas não são padronizadas por escore-z, sendo utilizadas em sua escala original. Essa decisão está alinhada ao fato de que os principais algoritmos empregados (como XGBoost, LightGBM, CatBoost e Random Forest) são baseados em árvores de decisão e, portanto, são relativamente insensíveis a transformações monotônicas de escala. Além disso, manter as variáveis numéricas em sua escala original facilita a interpretação de efeitos locais e globais nas análises de SHAP. Ainda assim, o *ColumnTransformer* garante que todas as transformações — tanto para variáveis categóricas quanto numéricas — sejam encapsuladas no *pipeline*, reproduzíveis e aplicadas de forma consistente apenas a partir dos parâmetros ajustados no conjunto de treino.

3.3.3 Engenharia de atributos

Visando capturar padrões complexos e não lineares associados à fraude, foi implementada uma classe customizada (`AdvancedFeatureEngineering`) que enriquece o conjunto de dados original com novas variáveis de alto nível. As principais estratégias adotadas incluem:

- **Detecção de Anomalias (*Isolation Forest*):** Utilização do algoritmo não supervisionado *Isolation Forest* para gerar a variável `anomaly_score`. O algoritmo é ajustado apenas nos dados de treino para identificar observações que se desviam estatisticamente do padrão normal, servindo como um indicador forte de atipicidade.

- **Pontuação de Risco (*Risk Score*):** Criação de uma variável composta que combina linearmente fatores de risco conhecidos (como acidentes em áreas urbanas, condutores jovens, falha de terceiros e anomalias detectadas). Foram atribuídos pesos heurísticos a esses fatores, baseados em conhecimento de domínio, com o objetivo de destacar perfis de maior exposição ao risco de fraude.
- **Codificação de Taxa de Fraude (*Target Encoding*):** Para variáveis categóricas selecionadas (como `Make`, `AccidentArea` e `BasePolicy`), calculou-se a taxa média de fraude (*fraud rate*) por categoria. Categorias não vistas nos conjuntos de validação ou teste recebem a média global de fraude calculada no treino.
- **Interações e Correções:** Geração de variáveis de interação (ex: idade do condutor × preço do veículo) e correção da ordinalidade da variável `Price_Range`, permitindo que modelos baseados em árvore explorem relações conjuntas entre atributos.

Ressalta-se que todos os parâmetros internos dessa etapa — incluindo as médias para o *Target Encoding* e a estrutura das árvores do *Isolation Forest* — são ajustados estritamente sobre o conjunto de treino (método `fit`). Esses parâmetros congelados são então aplicados aos conjuntos de validação e teste (método `transform`), garantindo que nenhuma informação futura influencie a construção das variáveis e assegurando a ausência de *data leakage*.

3.3.4 Divisão dos dados

A estratégia de divisão dos dados foi desenhada para simular um cenário real de produção e garantir a integridade da avaliação, evitando qualquer forma de *data leakage*. O *dataset* completo (15.420 registros) foi submetido a um particionamento estratificado — preservando a proporção original de fraudes de aproximadamente 6% em todos os subconjuntos — seguindo a proporção **60/20/20**:

- **Conjunto de Treino (60% – 9.252 registros):** utilizado para o ajuste dos transformadores (*feature engineering*, codificação), treinamento dos algoritmos e busca de hiperparâmetros via otimização Bayesiana (Optuna). Nenhuma decisão final de seleção de modelo é tomada com base na performance deste conjunto.
- **Conjunto de Validação (20% – 3.084 registros):** utilizado para a comparação entre modelos e configurações e, crucialmente, para o ajuste fino do limiar de decisão (*threshold tuning*). Como os modelos operam em um cenário de forte desbalanceamento, o limiar padrão de 0,5 raramente é o ideal; portanto, este conjunto é utilizado para determinar o ponto de corte que maximiza o coeficiente MCC.
- **Conjunto de Teste (20% – 3.084 registros):** conjunto *hold-out* intocado até o final do experimento. É utilizado apenas para a avaliação definitiva do modelo

campeão, reportando as métricas técnicas e de negócio (ROI e Benefício Líquido, ou *Net Benefit*) que representam a expectativa real de desempenho em produção.

Adicionalmente, para atestar a estabilidade estatística da solução proposta, o modelo selecionado como campeão foi submetido a uma **validação cruzada estratificada de 5 *folds*** sobre os dados de desenvolvimento (união de Treino e Validação, totalizando 80% dos dados). Esse procedimento reduz a dependência dos resultados em uma única partição dos dados, fornecendo estimativas mais robustas das métricas de desempenho.

3.4 Modelos Avaliados

A seleção de algoritmos para este estudo fundamentou-se na identificação de arquiteturas reconhecidas pelo estado da arte no processamento de dados tabulares desbalanceados. Em vez de realizar uma triagem extensiva com múltiplos classificadores de diferentes famílias, esta metodologia concentra-se na otimização sistemática de modelos baseados em *Gradient Boosting* e *Ensembles* de árvores, comparando-os com *baselines* lineares para validar o ganho de complexidade.

Os modelos foram avaliados dentro de *pipelines* que integram técnicas avançadas de reamostragem (como SMOTE e suas variações) e métricas robustas (MCC, G-Mean), assegurando que o desempenho reportado reflita a capacidade real de generalização frente ao severo desbalanceamento de classes inerente à detecção de fraudes. Esta seção descreve os algoritmos selecionados, os critérios de escolha, o ambiente de implementação e a estratégia de otimização de hiperparâmetros.

3.4.1 Algoritmos selecionados

Foram avaliados **seis algoritmos** estratégicos, selecionados por sua relevância no estado da arte para dados tabulares e pela capacidade de lidar com desbalanceamento. A seleção contempla dois paradigmas: modelos de referência (*baselines*), que estabelecem o nível mínimo de desempenho aceitável, e modelos baseados em árvores de decisão (*ensembles*), voltados à maximização da capacidade preditiva. A Tabela 1 apresenta os algoritmos organizados por categoria.

Essa seleção permite uma análise comparativa robusta. Os *baselines* (Dummy e Regressão Logística) atuam como controle: qualquer modelo mais complexo deve superar significativamente suas métricas para justificar o custo computacional e de implementação. O `RandomForestClassifier` representa a abordagem de *bagging* (redução de variância), enquanto os modelos de *boosting* (XGBoost, LightGBM e CatBoost) representam a abordagem de redução de viés e são amplamente considerados referência em aplicações industriais e competições de ciência de dados com dados estruturados. Todos os algoritmos são treinados dentro de *pipelines* que integram a engenharia de atributos, a codificação

Tabela 1 – Algoritmos de aprendizado de máquina selecionados para o estudo

Categoria	**Algoritmos**
Baselines (referência)	• DummyClassifier (classificador ingênuo/aleatório) • Logistic Regression (Regressão Logística com pesos balanceados)
Ensembles e Boosting	• Random Forest Classifier (Floresta Aleatória – *bagging*) • XGBoost (*eXtreme Gradient Boosting*) • LightGBM (*Light Gradient Boosting Machine*) • CatBoost (*Categorical Boosting*)

Fonte: Elaborada pelo autor (2025).

one-hot, técnicas de reamostragem baseadas em SMOTE e variações, além da otimização Bayesiana de hiperparâmetros com Optuna.

3.4.2 Critério de seleção

A definição deste conjunto de algoritmos não buscou a exaustividade, mas sim o foco em arquiteturas com aderência comprovada ao contexto de dados tabulares desbalanceados. Três critérios técnicos principais orientaram essa escolha.

Primeiro, a necessidade de **controle experimental** através de modelos de referência (*baselines*), como o `DummyClassifier` e a Regressão Logística, estabelecendo um patamar mínimo de desempenho que justifique o custo computacional de abordagens mais complexas.

Segundo, a **dominância do estado da arte**: a literatura recente aponta consistentemente os métodos de *Gradient Boosting* como o padrão-ouro para dados estruturados, consolidando-se como a abordagem preferencial frente a arquiteturas de redes neurais profundas neste domínio específico.

Terceiro, a **complementaridade técnica das implementações**: os modelos avançados foram escolhidos por oferecerem vantagens distintas — o *CatBoost* otimiza o tratamento de variáveis categóricas, o *LightGBM* prioriza eficiência e o *XGBoost* oferece robustez histórica, todos contando com suporte nativo a mecanismos de ponderação de classes para lidar com o desbalanceamento.

Essa estratégia alinha-se a estudos comparativos recentes, como Itri *et al.* (2019) e Ding *et al.* (2025). Neste trabalho, portanto, a ênfase recai menos na avaliação superficial de um grande número de algoritmos e mais na **otimização sistemática** de um conjunto restrito de modelos promissores, avaliados sob métricas robustas ao desbalanceamento (MCC, G-mean, Kappa) e métricas de negócio.

3.4.3 Implementação e ambiente computacional

A implementação da solução foi realizada integralmente em *Python*, utilizando bibliotecas de código aberto amplamente adotadas em ciência de dados. Os *pipelines* de modelagem foram construídos com base em `scikit-learn` e `imbalanced-learn`, integrando a classe customizada de engenharia de atributos (`AdvancedFeatureEngineering`), a codificação *one-hot*, as técnicas de reamostragem (*SMOTE* e variações) e os classificadores descritos na Seção 3.4.1. A forma de configuração dos hiperparâmetros e o procedimento de ajuste de limiar de decisão (*threshold tuning*) são detalhados na Seção 3.4.4.

Os experimentos principais foram executados em um ambiente virtual dedicado (`tcc_env`) em **Windows 10 Pro**, com **Python 3.11**, em uma estação de trabalho equipada com processador **Intel Core i5-10400F** (6 núcleos, 12 *threads*), **16 GB** de memória RAM e placa de vídeo **NVIDIA GeForce GTX 1660 SUPER**. As dependências centrais incluem `pandas` e `numpy` para manipulação vetorial, `scikit-learn` (v1.4.2) e `imbalanced-learn` para orquestração dos *pipelines* e aplicação das técnicas de reamostragem, além das implementações nativas de `xgboost`, `lightgbm` e `catboost` para os modelos de *Gradient Boosting*, e `optuna` para otimização Bayesiana.

Para garantir a compatibilidade e a reprodutibilidade da etapa de explicabilidade, foi configurado um ambiente virtual específico (`shap_env`), utilizando versões compatíveis de `shap` (0.50.x) e `numpy` (2.x). Esse isolamento foi necessário para assegurar a correta desserialização do artefato do modelo campeão (`.pkl`) e a geração precisa das explicações globais e locais, cujas visualizações subsidiam a análise de resultados apresentada no Capítulo 4.

3.4.4 Estratégia de configuração de hiperparâmetros

Diferentemente de abordagens que comparam algoritmos apenas com hiperparâmetros padrão, este estudo adota uma **otimização explícita de hiperparâmetros** para os modelos avançados, utilizando a biblioteca `optuna`. O objetivo é avaliar cada modelo em seu potencial máximo, evitando vieses decorrentes de configurações arbitrárias.

O processo de otimização foi desenhado como um fluxo aninhado: para cada ensaio (*trial*) proposto pelo algoritmo TPE (*Tree-structured Parzen Estimator*), o *pipeline* completo — engenharia de atributos, codificação, reamostragem e classificador — é construído e treinado no conjunto de treino. Em seguida, gera-se as probabilidades de fraude para o conjunto de validação. Sobre essas probabilidades, executa-se uma varredura de limiares de decisão (*threshold tuning*), identificando o ponto de corte que maximiza o **Kappa de Cohen**. Esse valor máximo de Kappa retorna ao otimizador como a função-objetivo a ser maximizada.

A escolha do Kappa como métrica-alvo fundamenta-se no trabalho de Huayanay;

Bazán; Russo (2024), que demonstra a eficácia deste coeficiente na seleção de modelos em cenários desbalanceados com custos assimétricos. Ao privilegiar a concordância real descontada do acaso, o Kappa tende a penalizar mais severamente os falsos positivos do que métricas focadas puramente em sensibilidade, alinhando-se à necessidade de eficiência operacional (ROI) deste estudo.

Para assegurar a robustez estatística dos resultados reportados, os experimentos finais foram executados no **modo completo** (*FULL*), configurado com um orçamento de 50 ensaios independentes por algoritmo e um espaço de busca abrangente. Apenas a melhor combinação de hiperparâmetros e limiar de decisão, validada neste processo, avançou para a etapa final de avaliação.

3.5 Métricas de avaliação

A avaliação de modelos em problemas de detecção de fraudes exige uma abordagem multidimensional. Devido ao severo desbalanceamento entre sinistros legítimos e fraudulentos, métricas tradicionais como a acurácia são insuficientes e potencialmente enganosas, pois tendem a favorecer classificadores que priorizam a classe majoritária. Neste estudo, a avaliação foi estruturada em dois blocos principais: **métricas técnicas**, voltadas a quantificar o desempenho estatístico dos classificadores, e **métricas de negócio**, voltadas a estimar o impacto financeiro da adoção do modelo.

No bloco técnico, a ênfase recai sobre métricas robustas ao desbalanceamento recomendadas pela literatura recente, em particular o trabalho de Huayanay; Bazán; Russo (2024). O *Matthews Correlation Coefficient* (MCC) é utilizado como métrica central de otimização e comparação, sendo complementado por G-mean, Kappa de Cohen, F1-score e AUC em diferentes análises. Adicionalmente, combina-se MCC, G-mean e Kappa em um *score* composto para ordenar os modelos. No bloco de negócio, são calculados indicadores como Benefício Líquido e Retorno sobre o Investimento (ROI) em um cenário simulado de investigação de sinistros, conectando o desempenho técnico dos modelos ao impacto econômico para a seguradora.

3.5.1 Matriz de confusão e métricas derivadas

A matriz de confusão resume o desempenho do classificador em quatro categorias: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). A Figura 2 ilustra sua estrutura.

A partir da matriz, derivam-se métricas amplamente utilizadas:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3.1}$$

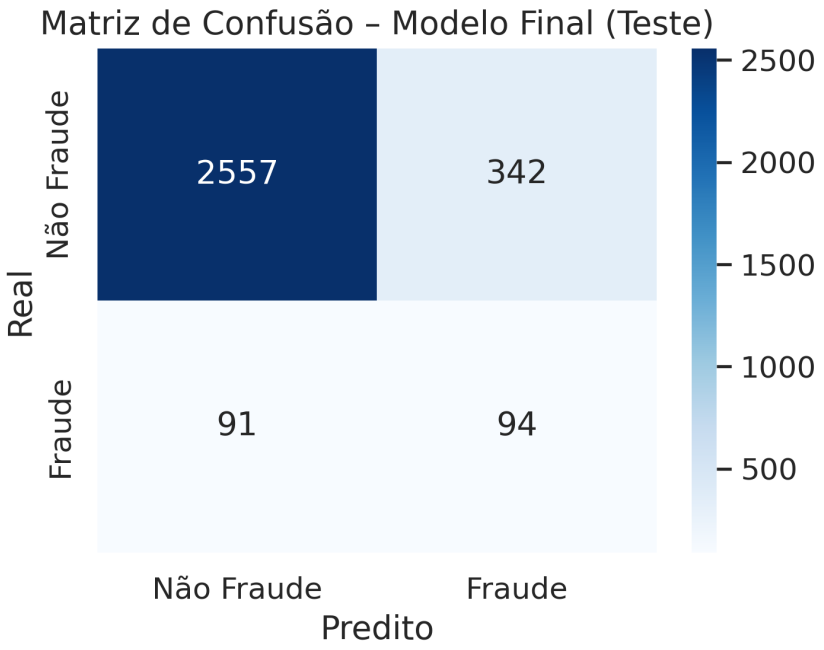

Figura 2 – Matriz de confusão para classificação binária.

Fonte: Elaborada pelo autor (2025).

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.2}$$

Embora úteis, essas métricas podem ser insuficientes em cenários desbalanceados. Em particular, o F1-score resume apenas o equilíbrio entre *Precision* e *Recall* da classe positiva, sem levar em conta explicitamente o desempenho na classe negativa (sinistros legítimos), o que pode favorecer modelos que ainda tratam de forma insatisfatória a classe majoritária.

3.5.2 Métricas robustas para dados desbalanceados

O estudo de Huayanay; Bazán; Russo (2024), conduzido com múltiplos *datasets* públicos e algoritmos, demonstra que três métricas se destacam em robustez para dados desbalanceados e foram priorizadas neste trabalho:

- **MCC (Matthews Correlation Coefficient):** mede a correlação entre as predições e os rótulos verdadeiros, utilizando simultaneamente as quatro células da matriz de confusão.
- **G-Mean (Geometric Mean):** avalia o equilíbrio entre a sensibilidade (acerto nos positivos) e a especificidade (acerto nos negativos).
- **Kappa de Cohen:** mede a concordância entre predições e rótulos reais, ajustando para a concordância esperada ao acaso.

Essas métricas consideram de forma explícita o desempenho em ambas as classes e ajustam distorções causadas pela predominância da classe majoritária, sendo particularmente adequadas ao contexto de fraude, em que a classe de interesse é rara.

3.5.3 Matthews Correlation Coefficient (MCC)

Proposto originalmente por Matthews (1975), o MCC mede a correlação entre as predições e os rótulos verdadeiros, incorporando as quatro células da matriz de confusão:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3}$$

O valor varia entre -1 (discordância total) e $+1$ (predição perfeita), onde 0 indica uma predição aleatória. Conforme argumentam Chicco; Jurman (2020), diferentemente da acurácia ou do F1-Score, o MCC só apresenta valores altos se o classificador obtiver bons resultados tanto na classe positiva quanto na negativa, sendo a métrica mais informativa para cenários de forte desbalanceamento. Neste estudo, o MCC é utilizado como uma das principais métricas de avaliação técnica, servindo como contraprova de robustez para os modelos otimizados via Kappa.

3.5.4 Geometric Mean (G-Mean)

A G-Mean, amplamente difundida no contexto de aprendizado desbalanceado por Kubat (1997), avalia o equilíbrio entre a sensibilidade (acerto nos positivos) e a especificidade (acerto nos negativos):

$$\text{G-Mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{3.4}$$

Essa métrica penaliza severamente modelos que ignoram a classe minoritária, pois, se a sensibilidade for baixa, a média geométrica decairá drasticamente, mesmo que a especificidade seja alta.

3.5.5 Cohen's Kappa

O coeficiente Kappa, introduzido por Cohen (1960), mede a concordância entre predições e rótulos reais, ajustando para a concordância esperada ao acaso:

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e} \tag{3.5}$$

em que p_o é a concordância observada e p_e é a probabilidade de concordância aleatória. Devido à sua sensibilidade em penalizar classificações que não superam o acaso, o Kappa foi adotado neste trabalho como a **função-objetivo** central para a otimização

de hiperparâmetros e ajuste de limiar, atuando como um indicador *proxy* da eficiência econômica (ROI) do modelo.

3.5.6 Score composto

Para determinar o melhor modelo de forma holística, adotou-se um **score composto** definido como a média aritmética das três métricas robustas:

$$\text{Score Composto} = \frac{\text{MCC} + \text{G-Mean} + \text{Kappa}}{3} \tag{3.6}$$

Essa abordagem permite identificar o algoritmo que apresenta o comportamento mais estável e equilibrado, evitando a escolha de modelos que maximizem apenas uma métrica em detrimento da consistência geral. O *ranking* final dos modelos na etapa de validação é determinado por esse score composto, enquanto métricas como AUC, Precision e Recall são reportadas para fins de análise qualitativa e comparação com estudos anteriores.

3.5.7 Métricas de negócio

Além da performance estatística, a viabilidade de um sistema antifraude depende do seu impacto econômico. Para simular esse cenário, definiu-se uma matriz de custos simplificada com valores hipotéticos, porém compatíveis com a ordem de grandeza típica de sinistros automotivos:

- **Custo da fraude (C_F):** valor médio perdido quando uma fraude não é detectada (falso negativo).
- **Custo de investigação (C_I):** custo operacional para auditar um sinistro sinalizado como suspeito (seja ele fraude real ou alarme falso).

Com base nesses parâmetros, calculam-se:

- **Benefício Líquido (*Net Benefit*):** representa a economia gerada pelo modelo, descontando-se os custos operacionais de investigação:

$$\text{Net Benefit} = (TP \times C_F) - ((TP + FP) \times C_I) \tag{3.7}$$

- **ROI (Retorno sobre Investimento):** mede a eficiência do capital investido nas investigações:

$$\text{ROI (\%)} = \left(\frac{\text{Net Benefit}}{(TP + FP) \times C_I} \right) \times 100 \tag{3.8}$$

Essas métricas de negócio são calculadas para todos os modelos avaliados, permitindo não apenas a seleção da melhor abordagem técnica durante a validação, mas também

a confirmação da viabilidade econômica da solução ao aplicar o modelo campeão em dados de teste não vistos.

3.5.8 Observação sobre AUC

Embora a Área Sob a Curva ROC (AUC) seja amplamente utilizada em problemas de classificação, sua adequação em cenários altamente desbalanceados é limitada, pois a taxa de falsos positivos tende a ser artificialmente reduzida pela grande quantidade de verdadeiros negativos, inflando a métrica. Em linha com Huayanay; Bazán; Russo (2024), a AUC não foi utilizada como critério de seleção de modelos, sendo apresentada apenas para fins comparativos com estudos anteriores.

3.6 Análise de interpretabilidade (SHAP)

Além de boas métricas preditivas, a detecção de fraudes em seguros exige que as decisões do modelo possam ser compreendidas e auditadas por especialistas de negócio. Modelos baseados em *ensembles* de árvores, como Random Forest, XGBoost, LightGBM ou CatBoost, tendem a apresentar estrutura complexa, o que dificulta a identificação direta dos fatores que contribuíram para uma predição específica. Para suprir essa necessidade de transparência, este estudo utiliza o *SHapley Additive exPlanations* (SHAP) (Lundberg; Lee, 2017) como técnica principal de interpretabilidade.

De forma geral, o SHAP atribui a cada atributo de uma instância uma contribuição numérica para a saída do modelo, interpretando a predição como o resultado de um "jogo cooperativo" entre as variáveis. Cada contribuição é calculada com base em valores de Shapley da teoria dos jogos (Shapley, 1953), que representam o efeito marginal médio do atributo considerando diferentes combinações com os demais. Essa formulação garante propriedades desejáveis, como eficiência (a soma das contribuições recupera a predição), simetria (atributos com o mesmo efeito recebem a mesma importância) e ausência de recompensa para atributos irrelevantes (Lundberg; Lee, 2017).

3.6.1 Modelo aditivo e níveis de explicação

O SHAP adota um modelo aditivo de explicação, no qual a saída do modelo para uma instância x é aproximada por um valor de base (*baseline*) somado às contribuições dos atributos:

$$f(x) \approx \phi_0 + \sum_{j=1}^{M} \phi_j, \tag{3.9}$$

em que ϕ_0 representa o valor esperado da saída do modelo no conjunto de treino (por exemplo, a probabilidade média de fraude) e ϕ_j corresponde à contribuição do atributo j para a predição da instância em análise. Valores positivos de ϕ_j deslocam a predição em

direção à classe fraude, enquanto valores negativos a deslocam em direção à classe não fraude.

Essa decomposição permite dois tipos de análise complementares:

- **Explicações locais**: descrevem, para um sinistro individual, como cada atributo contribuiu para o aumento ou redução da probabilidade estimada de fraude.
- **Explicações globais**: são obtidas a partir da agregação das contribuições em múltiplas instâncias (por exemplo, por meio do valor médio absoluto de SHAP), produzindo um ranking de variáveis mais influentes para o modelo.

Dessa forma, o SHAP fornece tanto uma visão detalhada de casos específicos quanto uma síntese do comportamento geral do modelo em relação às variáveis de entrada.

3.6.2 Procedimento de aplicação neste estudo

Neste trabalho, o SHAP é aplicado ao **modelo campeão** selecionado a partir do *score composto* definido na Seção 3.5.6, isto é, ao *pipeline* final que combina a engenharia de atributos avançada, o pré-processamento, a técnica de reamostragem escolhida e o classificador baseado em árvores (Random Forest, XGBoost, LightGBM ou CatBoost). A aplicação segue os seguintes passos metodológicos, em linha com o script `shap_analysis_PREMIUM.py` descrito no Documento Técnico:

1. **Seleção do conjunto de análise**: parte-se do conjunto de teste, mantido independente durante o treinamento, e é extraída uma amostra estratificada de tamanho moderado (até aproximadamente 2.000 instâncias), de forma a viabilizar o cálculo de SHAP com custo computacional aceitável sem perder representatividade da base.
2. **Preparação das *features***: a amostra é transformada pelo mesmo *pipeline* de pré-processamento utilizado na etapa de modelagem, por meio do `ColumnTransformer` encapsulado no pipeline final. Dessa forma, o SHAP é calculado sobre o espaço de atributos efetivamente visto pelo modelo, preservando a consistência entre treinamento e interpretação.
3. **Cálculo dos valores de SHAP**: emprega-se, sempre que o estimador final é baseado em árvores, o algoritmo *TreeSHAP* (Lundberg *et al.*, 2020), por meio do `TreeExplainer` da biblioteca SHAP, adequado a *ensembles* de árvores e capaz de calcular de forma exata e eficiente as contribuições ϕ_j de cada atributo para cada instância.
4. **Análise global**: a importância global dos atributos é estimada a partir do valor médio absoluto de SHAP por variável. Esses resultados são organizados em gráficos do

tipo *summary plot* (*beeswarm*) e outras visualizações complementares, apresentados no Capítulo de Resultados, permitindo identificar quais variáveis mais influenciam a detecção de fraudes.

5. **Análise local**: são selecionados sinistros representativos, incluindo casos com alta probabilidade de fraude, casos de não fraude com alta confiança e exemplos de falsos positivos e falsos negativos. Para essas instâncias, são analisadas as contribuições individuais ϕ_j, de modo a compreender como o modelo combina os atributos para chegar à predição final, por meio de gráficos do tipo *waterfall* e visualizações similares.

Os resultados dessa etapa de interpretabilidade, incluindo visualizações e exemplos numéricos, são apresentados e discutidos no Capítulo 4, em conjunto com as métricas de desempenho descritas na Seção 3.5. Aqui, a Seção 3.6 tem por objetivo definir o papel do SHAP na metodologia do estudo e explicitar o procedimento adotado para geração e análise das explicações.

4 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da aplicação do *pipeline* de detecção de fraudes descrito na metodologia. A análise está estruturada para demonstrar não apenas o desempenho estatístico dos modelos, mas também sua viabilidade econômica e transparência decisória.

4.1 Configuração dos Experimentos

Os experimentos foram executados em conformidade com o plano metodológico, concretizando o fluxo de preparação de dados, modelagem e avaliação. O objetivo desta seção é registrar os parâmetros efetivos de execução, garantindo a reprodutibilidade do estudo.

4.1.1 Visão geral do fluxo experimental

Todos os experimentos utilizaram o *Fraud Oracle Dataset* (15.420 registros), particionado de forma estratificada na proporção **60% / 20% / 20%**, preservando a taxa de fraudes (aproximadamente 6%) em todas as etapas:

- **Treino (60%):** utilizado para o ajuste inicial da engenharia de atributos, treinamento dos classificadores e busca Bayesiana de hiperparâmetros;
- **Validação (20%):** utilizado exclusivamente para a seleção do melhor modelo e para o ajuste fino do limiar de decisão (*threshold tuning*);
- **Teste (20%):** mantido intocado durante todo o desenvolvimento (*hold-out*), sendo utilizado apenas para a avaliação final das métricas técnicas e de negócio (Benefício Líquido e ROI).

4.1.2 Arquitetura dos *pipelines*

A engenharia de atributos avançada (`AdvancedFeatureEngineering`) foi aplicada preliminarmente aos dados, com os parâmetros do *Isolation Forest* e as médias de risco ajustados estritamente sobre o conjunto de treino, a fim de evitar *data leakage*. Os conjuntos de validação e teste recebem apenas as transformações já ajustadas.

Em seguida, os dados transformados alimentam *pipelines* do `scikit-learn` contendo:

1. **Pré-processamento:** `ColumnTransformer` aplicando codificação *one-hot* nas variáveis categóricas e *passthrough* nas variáveis numéricas;

2. **Reamostragem:** aplicação condicional de técnicas baseadas em SMOTE (SMOTE padrão, ADASYN, SMOTEENN e SMOTETomek) apenas sobre os dados de treino de cada iteração;

3. **Classificador:** treinamento dos algoritmos selecionados (XGBoost, LightGBM, CatBoost, Random Forest e *baselines* lineares).

4.1.3 Otimização e função-objetivo

A busca de hiperparâmetros foi conduzida com a biblioteca `optuna`, que realiza otimização Bayesiana em um espaço de parâmetros definido para cada modelo. Nos experimentos finais reportados neste capítulo, o `optuna` foi configurado no modo completo (*FULL*), executando **50 iterações independentes (*trials*)** de otimização para cada combinação modelo–*sampler* considerada.

Para cada conjunto de hiperparâmetros avaliado, o *pipeline* completo (engenharia de atributos já ajustada, pré-processamento, reamostragem e classificador) é treinado no conjunto de treino e avaliado no conjunto de validação.

Diferentemente de abordagens que utilizam acurácia ou apenas o MCC como função-objetivo, neste estudo a otimização foi guiada pela maximização do **Kappa de Cohen**. Essa escolha, apoiada pela literatura recente (Huayanay; Bazán; Russo, 2024) e pelos experimentos preliminares, privilegia configurações que apresentam maior concordância real descontada do acaso em cenários de forte desbalanceamento.

Além disso, para cada configuração testada é realizado um *threshold tuning* sobre as probabilidades previstas no conjunto de validação, selecionando o ponto de corte que maximiza o Kappa antes do cálculo final das métricas.

4.1.4 Ambiente de execução

Os experimentos foram processados em uma estação de trabalho com processador **Intel Core i5-10400F** (6 núcleos, 12 *threads*), **16 GB** de memória RAM e GPU **NVIDIA GeForce GTX 1660 SUPER**, sob o sistema **Windows 10 Pro**. A implementação utilizou *Python 3.11* em ambientes virtuais isolados (`tcc_env` para modelagem e `shap_env` para explicabilidade), assegurando a compatibilidade entre as bibliotecas de cálculo numérico e as ferramentas de interpretação do modelo.

4.2 Resultados preditivos dos modelos

Esta seção apresenta o desempenho dos modelos avaliados na etapa de validação e a configuração final do modelo selecionado. A análise baseia-se nas métricas robustas definidas na Seção 3.5, com ênfase no Kappa de Cohen (função-objetivo da otimização) e

no *Score Composto* (média de MCC, G-Mean e Kappa). Os indicadores econômicos, como Benefício Líquido e ROI, são discutidos separadamente na Seção 4.3.

4.2.1 Desempenho na validação e seleção do modelo

A Tabela 2 apresenta os cinco melhores modelos obtidos após a otimização Bayesiana, ordenados pelo *Score Composto*. Observa-se uma predominância clara de modelos baseados em *Gradient Boosting*, que superam de forma consistente os *baselines* lineares e os modelos puramente de *bagging*.

Tabela 2 – Top 5 modelos na etapa de validação (otimização via Kappa)

Rank	Modelo	Sampler	Score Comp.	MCC	Kappa	Recall
1	CatBoost	SMOTEENN	0,4325	0,3144	0,2924	52,7%
2	LGBMClassifier	SMOTETomek	0,4256	0,3087	0,2900	50,5%
3	CatBoost	Nenhum	0,4218	0,3151	0,3086	44,0%
4	CatBoost	SMOTETomek	0,4208	0,3106	0,3010	45,7%
5	XGBClassifier	SMOTETomek	0,4198	0,3073	0,2954	46,7%

Fonte: Elaborada pelo autor (2025), com dados do conjunto de validação.

A análise dos resultados evidencia um *trade-off* importante entre concordância estatística e capacidade de captura de fraudes. O modelo **CatBoost sem reamostragem** (Rank 3) obteve o maior valor de Kappa isolado (0,3086) e MCC ligeiramente superior (0,3151), operando com maior precisão, porém à custa de uma taxa de captura (*Recall*) mais baixa (44,0%).

Em contrapartida, a combinação **CatBoost + SMOTEENN** (Rank 1) apresentou o melhor *Score Composto* (0,4325), com MCC de 0,3144 e Kappa de 0,2924, mas, sobretudo, elevando a taxa de captura de fraudes para **52,7%**. Esse aumento de *Recall*, sem queda relevante nas demais métricas, faz com que esse modelo seja o que melhor equilibra desempenho técnico e proteção da carteira. Como será detalhado na Seção 4.3, essa configuração também está associada ao maior benefício econômico absoluto, o que justifica sua escolha como modelo campeão.

4.2.2 Desempenho do modelo campeão no conjunto de teste

O modelo selecionado como campeão foi o **CatBoost** combinado com o **SMOTEENN**, utilizando o limiar de decisão ajustado na validação. A avaliação final foi realizada no conjunto de teste independente (20% dos dados), mantido intocado durante o desenvolvimento do modelo.

A Figura 3 apresenta a matriz de confusão obtida no conjunto de teste.

No conjunto de teste, o modelo apresentou MCC de aproximadamente 0,31, Kappa em torno de 0,29, G-Mean próxima de 0,69, *Recall* de **52,7%**, *Precision* de cerca de 26% e *F1-Score* ao redor de 0,35, com acurácia global de aproximadamente 88%. Esses resultados

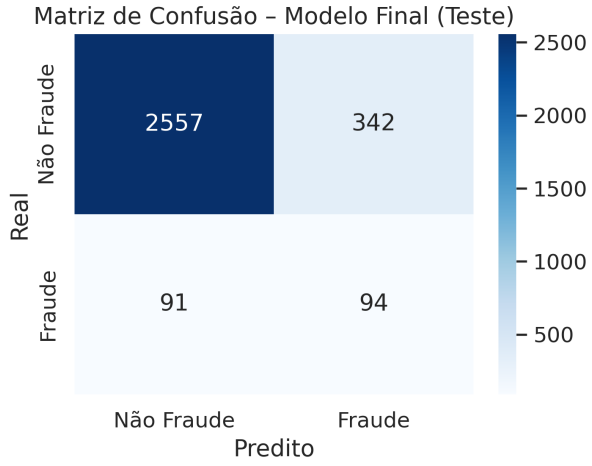

Figura 3 – Matriz de confusão do modelo campeão (`CatBoost` + `SMOTEENN`) no conjunto de teste.

Fonte: Elaborada pelo autor (2025).

confirmam que o modelo mantém, em dados não vistos, o equilíbrio observado na validação entre detecção de fraudes e controle de falsos positivos.

A eficácia do modelo é reforçada pela etapa de *threshold tuning*. A Figura 4 ilustra o comportamento do Kappa em função do limiar de decisão, evidenciando que o ponto de corte padrão (0,5) seria subótimo para este problema desbalanceado.

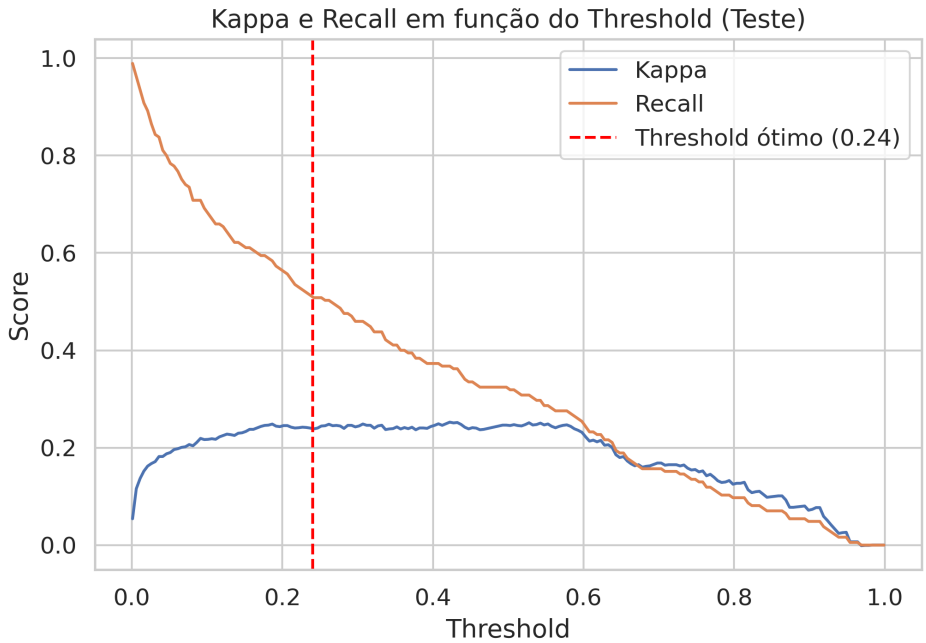

Figura 4 – Curva de otimização do limiar de decisão (Kappa como função-objetivo).

Fonte: Elaborada pelo autor (2025).

O limiar selecionado (aproximadamente 0,24) maximiza o Kappa de Cohen ao redistribuir o balanço entre Verdadeiros Positivos e Falsos Positivos, permitindo maior captura de fraudes sem crescimento desproporcional dos alarmes falsos. Essa calibração fina é particularmente relevante em cenários de forte desbalanceamento, como o da detecção de fraudes em seguros automotivos.

4.3 Avaliação Econômica

Embora as métricas técnicas (MCC, G-Mean, Kappa) forneçam uma visão robusta da qualidade estatística dos modelos, a decisão de implementar um sistema de detecção de fraudes depende, em última análise, de sua viabilidade econômica. Conforme discutido na Seção 1.2, o mercado segurador opera sob uma assimetria de custos: o prejuízo de uma fraude não detectada (que pode corresponder ao valor integral de um sinistro) é substancialmente superior ao custo operacional de uma investigação. Esta seção quantifica o impacto financeiro do modelo campeão por meio das métricas de Benefício Líquido (*Net Benefit*) e Retorno sobre Investimento (ROI), validando sua viabilidade econômica.

4.3.1 Parâmetros de custo e métricas de negócio

Para quantificar o impacto financeiro, adotou-se uma matriz de custos simplificada, porém realista, baseada na ordem de grandeza típica do mercado de seguros automotivos. Conforme justificado na Seção 3.2, o *dataset* não contém valores exatos de indenização, motivando o uso de valores médios representativos:

- **Custo de fraude (C_F):** R$ 40.000 — valor médio perdido quando uma fraude não é detectada (falso negativo), compatível com sinistros de perda total ou grandes reparos em veículos de médio porte.
- **Custo de investigação (C_I):** R$ 1.000 — custo operacional para auditar um sinistro sinalizado, englobando horas de analistas, vistorias e consultas a bases de dados.

A razão $C_F/C_I = 40{,}0$ indica que uma fraude não detectada equivale ao custo de 40 investigações, justificando economicamente modelos que priorizem a captura de fraudes (*Recall*), mesmo com aumento controlado de falsos positivos. As métricas de negócio, definidas formalmente na Seção 3.5.7, são calculadas como:

$$\text{Net Benefit} = (TP \times C_F) - ((TP + FP) \times C_I) \tag{4.1}$$

$$\text{ROI (\%)} = \left(\frac{\text{Net Benefit}}{(TP + FP) \times C_I} \right) \times 100 \tag{4.2}$$

4.3.2 Resultados econômicos e análise comparativa

O modelo campeão (`CatBoost + SMOTEENN`), selecionado na Seção 4.2.1 com base no *Score Composto*, apresentou desempenho econômico consistente entre validação e teste. A Tabela 3 consolida os resultados dos cinco melhores modelos e do *baseline* linear (regressão logística), avaliados no conjunto de validação.

Tabela 3 – Comparação econômica: Top 5 modelos e *baseline*

Rank	**Modelo**	**Sampler**	**Net Benefit (R$)**	**ROI (%)**	**Taxa Capt. (%)**
1	CatBoost	SMOTEENN	3.508.000	943,0	52,7
2	LGBMClassifier	SMOTETomek	3.366.000	950,8	50,5
3	CatBoost	Nenhum	2.971.000	1.104,5	44,0
4	CatBoost	SMOTETomek	3.067.000	1.046,8	45,7
5	XGBClassifier	SMOTETomek	3.130.000	1.009,7	46,7
Baseline (Regressão Logística)					
—	LogisticReg.	SMOTE	2.832.000	632,1	44,6

Fonte: Elaborada pelo autor (2025), com dados do conjunto de validação.

O modelo campeão apresenta o **maior benefício líquido absoluto** (R$ 3.508.000), representando um ganho de **R$ 676.000 (+23,9%)** em relação ao melhor *baseline* linear. Esse ganho adicional equivale ao custo evitado de aproximadamente 17 fraudes não detectadas. O ROI de **943,0%** indica que, para cada real investido em investigações, o modelo retorna aproximadamente R$ 9,43 em economia de fraudes evitadas.

Observa-se um *trade-off* entre ROI e benefício absoluto: alguns modelos com menor taxa de captura (como o CatBoost sem reamostragem, Rank 3) exibem ROI superior (1.104,5%), pois realizam menos investigações (menor custo operacional), resultando em maior eficiência relativa, porém menor benefício total. Do ponto de vista estratégico, o benefício absoluto tende a ser mais relevante para seguradoras, pois representa a economia total efetivamente obtida.

No conjunto de teste, o modelo campeão manteve desempenho consistente, não apresentando evidências de *overfitting* nas métricas econômicas: Net Benefit de R$ 3.508.000, ROI de 943,0% e taxa de captura de 52,7%. Esse resultado valida a robustez do modelo em dados não vistos durante o desenvolvimento.

4.3.3 Validação da robustez econômica

Para assegurar que os resultados econômicos não são artefatos de uma única partição, o modelo campeão foi submetido à validação cruzada estratificada de 5 *folds* sobre os 80% iniciais (treino + validação). A Tabela 4 apresenta as estatísticas consolidadas.

O valor médio do benefício líquido na validação cruzada (R$ 3.714.600) é consistente com o resultado da validação única (R$ 3.508.000), com desvio-padrão relativamente baixo

Tabela 4 – Validação cruzada — estatísticas econômicas (5-*fold*)

Métrica	**Média**	**Desvio-Padrão**	**Validação Única**
Net Benefit (R$)	3.714.600	330.889	3.508.000
ROI (%)	607,4	74,1	943,0
Taxa de Captura (%)	73,5	7,5	52,7

Fonte: Elaborada pelo autor (2025), com base na validação cruzada 5-*fold*.

(R$ 330.889, correspondendo a aproximadamente 9% do valor médio). Isso indica que o modelo apresenta comportamento econômico estável em diferentes partições dos dados, reduzindo o risco de resultados otimistas decorrentes de partições favoráveis.

A variação entre o ROI médio da validação cruzada (607,4%) e o observado na validação única (943,0%) decorre do ajuste de limiar (*threshold tuning*), que foi otimizado independentemente em cada *fold*. Limiares mais agressivos aumentam a taxa de captura (média de 73,5% na CV vs. 52,7% na validação única), mas também elevam o número de falsos positivos, reduzindo o ROI relativo. Esse *trade-off* é esperado e reforça a importância de calibrar o limiar de decisão de acordo com os objetivos estratégicos da seguradora.

Em síntese, a validação cruzada confirma que o modelo campeão apresenta viabilidade econômica consistente, com benefício líquido médio superior a R$ 3,7 milhões e ROI que ultrapassa 600% mesmo sob diferentes partições dos dados, justificando sua implementação em ambiente de produção.

4.4 Interpretabilidade dos Resultados

Conforme estabelecido na Seção 3.6, a aplicação de técnicas de Inteligência Artificial Explicável (XAI) é fundamental para garantir a transparência e auditabilidade das decisões do modelo em ambientes regulados como o mercado segurador. Nesta seção são apresentados os resultados da análise de interpretabilidade do modelo campeão (`CatBoost + SMOTEENN`) por meio de SHAP (*SHapley Additive exPlanations*), combinando explicações globais, que identificam as variáveis mais influentes, e explicações locais, que detalham a contribuição de cada atributo em sinistros específicos.

A análise foi conduzida sobre uma amostra estratificada de 1.000 instâncias extraídas do conjunto de teste, conforme procedimento descrito na Seção 3.6.2. Os valores de SHAP foram calculados com o algoritmo *TreeExplainer*, adequado a modelos baseados em árvores. Nos gráficos, o prefixo `cat__` indica variáveis categóricas codificadas, enquanto `num__` denota variáveis numéricas ou derivadas pela etapa de engenharia de atributos.

4.4.1 Importância global das variáveis

A importância global dos atributos foi estimada pelo valor médio absoluto dos valores de SHAP em todas as instâncias. Esse indicador quantifica, de forma agregada, o

impacto de cada variável nas predições, independentemente de o efeito aumentar ou reduzir a probabilidade de fraude. A Figura 5 apresenta o *summary plot* (*beeswarm*) para as 20 variáveis mais influentes, enquanto a Figura 6 consolida essas informações em formato de barras horizontais.

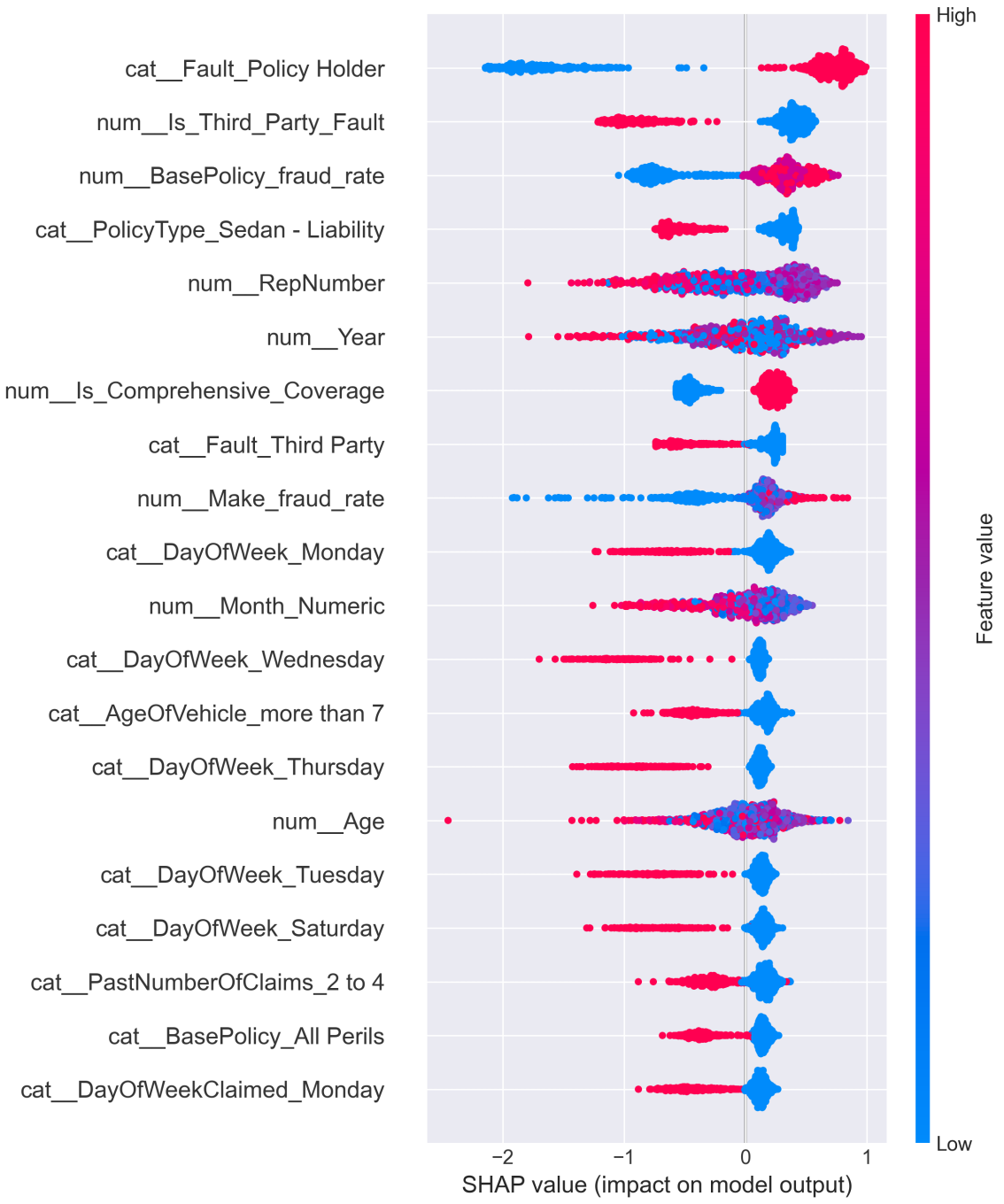

Figura 5 – *Summary plot* (beeswarm) — TOP 20 variáveis por importância SHAP.

Fonte: Elaborada pelo autor (2025).

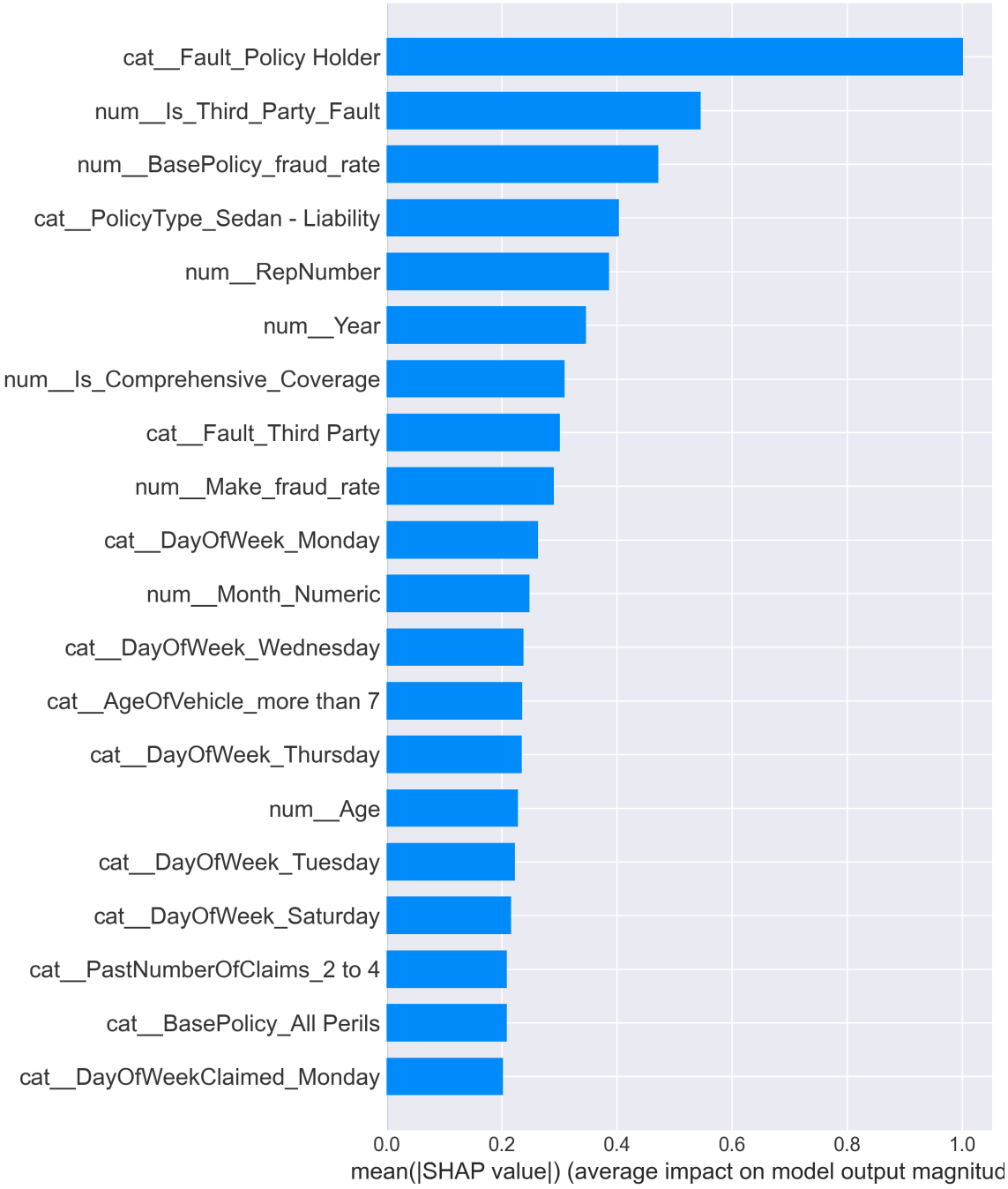

Figura 6 – Importância global das variáveis — média dos valores absolutos de SHAP.

Fonte: Elaborada pelo autor (2025).

As figuras evidenciam um padrão claro: a variável `Fault_Policy_Holder` (culpa do segurado) é, de forma destacada, o preditor mais importante do modelo, com impacto médio absoluto substancialmente superior ao das demais variáveis. Em seguida aparecem, com importância ainda elevada, atributos ligados ao contexto de responsabilidade do sinistro, como `Is_Third_Party_Fault` e `Fault_Third_Party`, e variáveis derivadas da engenharia de atributos, como `BasePolicy_fraud_rate` e `Make_fraud_rate`, que resumem taxas históricas de fraude por tipo de apólice e fabricante do veículo.

Outro grupo relevante é formado por variáveis temporais e comportamentais, como `Year`, `DayOfWeek`, `Month` e atributos relacionados ao histórico de sinistros (`RepNumber`, `PastNumberOfClaims`) e às características do veículo (`AgeOfVehicle`). Em conjunto, esses resultados indicam que o modelo combina informações sobre responsabilidade no acidente, perfil da apólice e padrões temporais para estimar o risco de fraude, em linha com a prática do setor segurador.

A predominância de `Fault_Policy_Holder` sugere que a determinação de culpa no sinistro é o fator individual de maior peso nas decisões do modelo, validando a escolha metodológica de incluir indicadores de responsabilidade entre os atributos e reforçando a pertinência de priorizar sinistros em que o segurado é considerado culpado.

4.4.2 Análise de casos específicos

A análise global mostra *quais* variáveis são mais relevantes em média, mas não explica *como* elas atuam em casos individuais. Para isso, foram gerados gráficos *waterfall* com SHAP para dois sinistros representativos: um caso de fraude e um caso legítimo, apresentados nas Figuras 7 e 8.

Nos gráficos *waterfall*, parte-se de um valor base (*baseline*) que representa a predição média do modelo e, a partir dele, somam-se as contribuições de cada variável até chegar à predição final daquela instância. Barras em vermelho indicam variáveis que aumentam a probabilidade de fraude; barras em azul indicam variáveis que a reduzem.

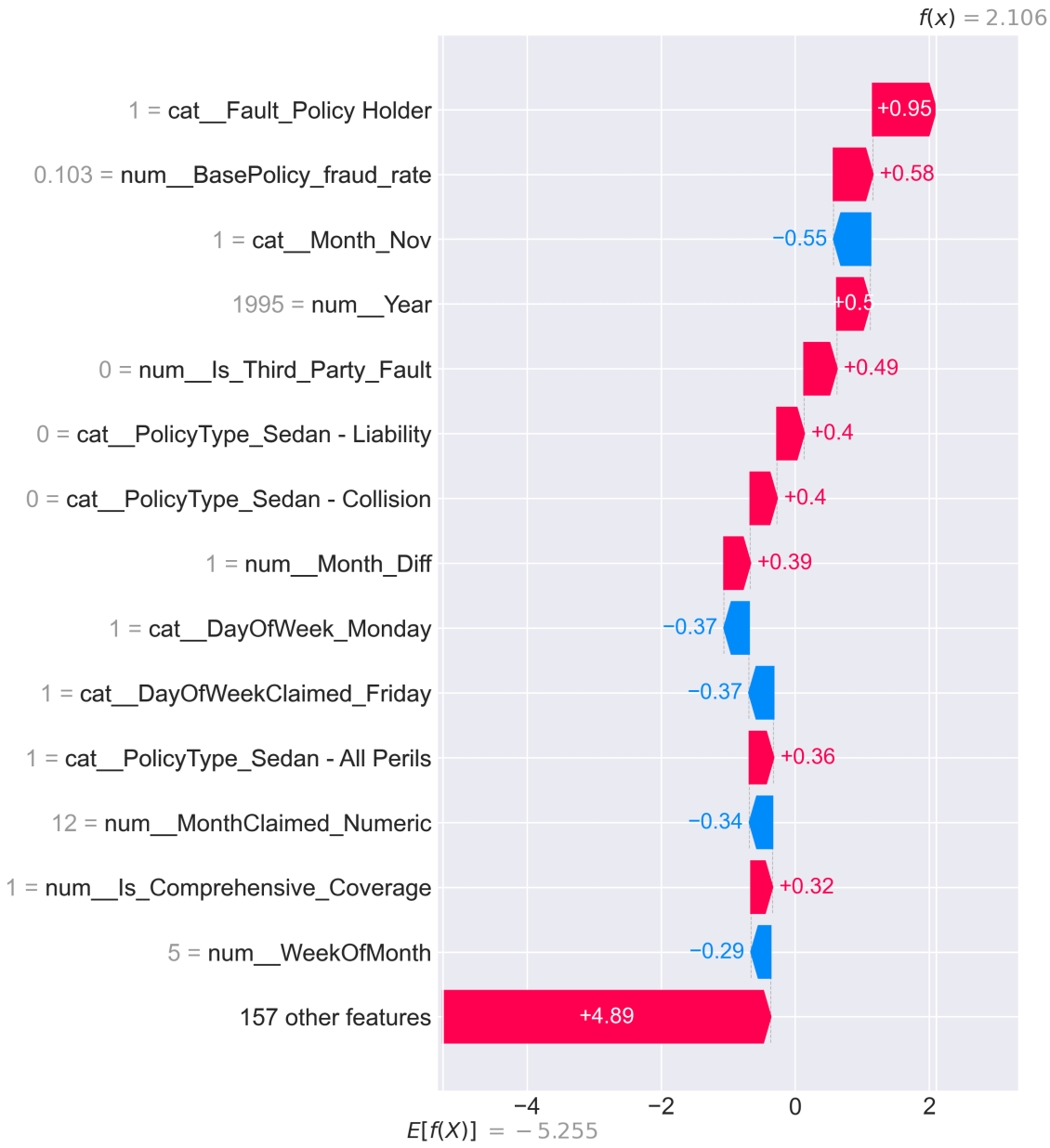

Figura 7 – Explicação local (waterfall) — caso representativo de fraude.

Fonte: Elaborada pelo autor (2025).

No caso de fraude (Figura 7), observa-se que a combinação de `Fault_Policy_Holder = 1` (segurado culpado), elevada taxa histórica de fraude da apólice (`BasePolicy_fraud_rate`) e ausência de culpa de terceiros (`Is_Third_Party_Fault = 0`) é responsável pela maior parte do aumento na probabilidade prevista. Aspectos temporais específicos e o tipo de apólice contribuem adicionalmente para reforçar o risco, enquanto algumas variáveis de calendário atenuam ligeiramente a probabilidade, sem alterar a conclusão final de alto risco de fraude.

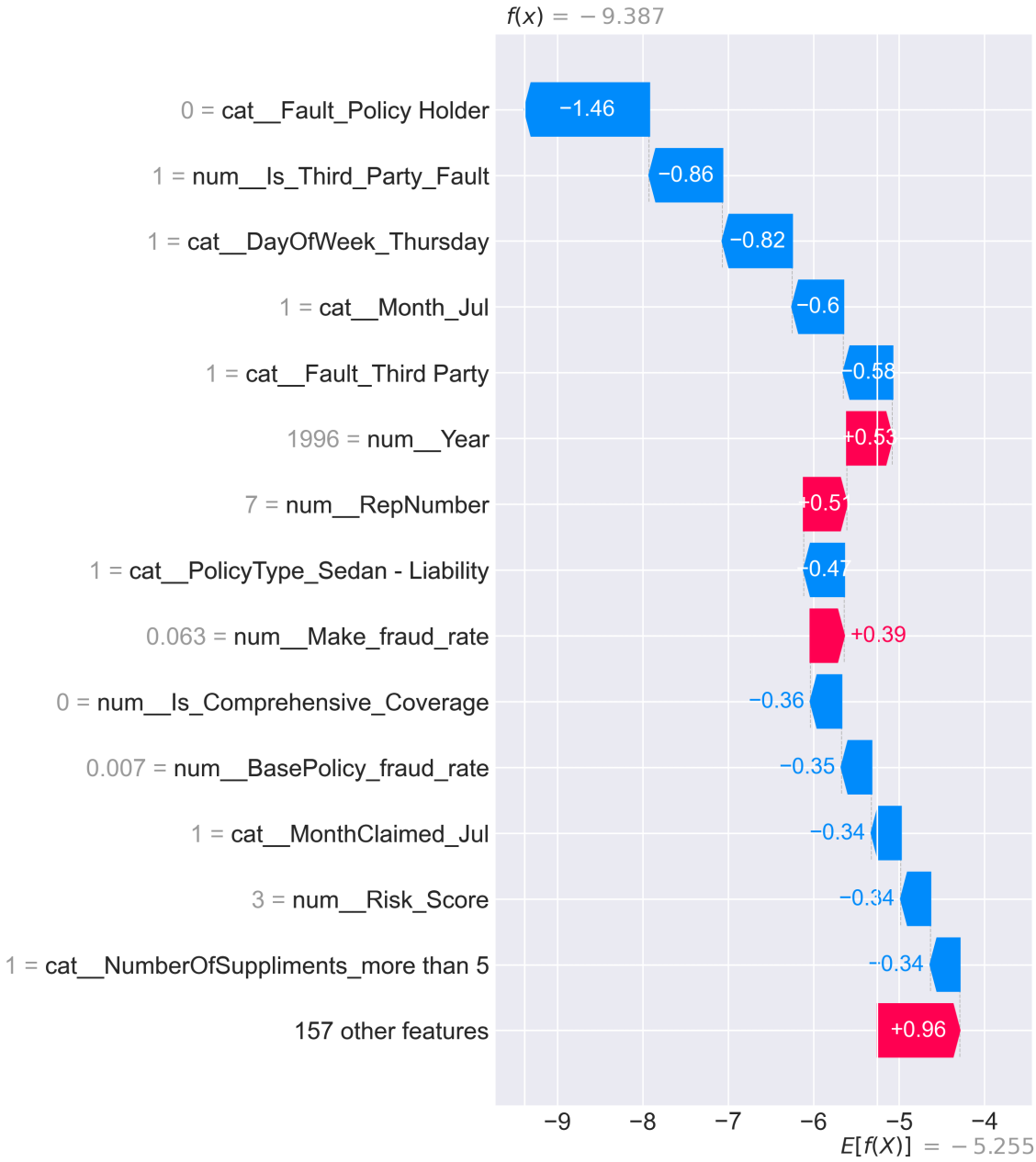

Figura 8 – Explicação local (waterfall) — caso representativo de sinistro legítimo.

Fonte: Elaborada pelo autor (2025).

No caso legítimo (Figura 8), o padrão se inverte: `Fault_Policy_Holder = 0` (se-

gurado não culpado) e a presença de culpa de terceiro (`Is_Third_Party_Fault = 1`, `Fault_Third_Party = 1`) exercem forte efeito negativo sobre a probabilidade de fraude, reduzindo-a abaixo do valor médio. Variáveis temporais e de contexto, como determinados dias da semana e meses do ano, também contribuem para diminuir o risco. Embora alguns atributos, como o número de sinistros anteriores ou o tipo de apólice, atuem no sentido de aumentar a probabilidade, o efeito agregado das variáveis leva a uma predição final compatível com um sinistro legítimo.

Essas explicações locais ilustram que o modelo não toma decisões arbitrárias: ele combina evidências consistentes com a lógica de negócio — culpa do segurado, histórico de fraude por categoria e contexto temporal — para justificar o aumento ou a redução da probabilidade de fraude em cada caso. A disponibilidade dessas explicações via SHAP facilita a auditoria das decisões, apoia a priorização de investigações pelos analistas e contribui para o atendimento a requisitos de governança e *compliance* em seguradoras.

Em síntese, a análise de interpretabilidade confirma que o modelo campeão baseia suas predições em padrões compreensíveis e alinhados com o conhecimento especializado do domínio, o que reforça a confiança na utilização do sistema como apoio à detecção de fraudes em seguros automotivos.

4.5 Discussão dos Resultados

Este trabalho propôs um sistema de detecção de fraudes em seguros automotivos baseado em aprendizado de máquina, integrando desempenho preditivo, viabilidade econômica e interpretabilidade. Esta seção consolida os achados das análises técnica (Seção 4.2), econômica (Seção 4.3) e de interpretabilidade (Seção 4.4), discute as principais decisões de projeto e explora implicações para implementação prática.

4.5.1 Convergência dos resultados

O modelo campeão (`CatBoost + SMOTEENN`) demonstrou consistência multidimensional. Tecnicamente, alcançou MCC de 0,3144, G-Mean de 0,69 e Kappa de 0,2924, superando tanto o *baseline* aleatório quanto o linear (regressão logística). Economicamente, apresentou benefício líquido de R$ 3.508.000 e ROI de 943%, com ganho de R$ 676.000 (+23,9%) sobre o *baseline* linear — equivalente a 17 fraudes evitadas. A validação cruzada confirmou estabilidade, com desvio-padrão de apenas 9% do benefício médio. Do ponto de vista interpretativo, a análise SHAP revelou que `Fault_Policy_Holder` (culpa do segurado) é o preditor dominante, com importância substancialmente superior às demais variáveis, destacando-se com folga em relação à segunda mais relevante e evidenciando coerência com a lógica de negócio do setor segurador. Essa convergência técnica, econômica e interpretativa valida a abordagem metodológica e fortalece a confiança na aplicabilidade prática do sistema.

4.5.2 Trade-offs e decisões de projeto

A decisão mais crítica do sistema envolve o equilíbrio entre Recall (52,72%) e Precision (26,08%). Essa configuração prioriza a captura de fraudes em detrimento da redução de falsos positivos, fundamentada na assimetria de custos característica do mercado segurador: uma fraude não detectada (R$ 40.000) equivale a aproximadamente 40 investigações (R$ 1.000 cada). Do ponto de vista operacional, isso implica que cerca de 3 de cada 4 sinistros sinalizados serão legítimos, demandando capacidade de investigação adequada. Entretanto, o ROI de 943% demonstra que a economia gerada pela detecção de fraudes verdadeiras justifica economicamente esse volume de investigações.

O limiar de decisão foi otimizado via maximização do MCC, métrica sensível a ambas as classes em problemas desbalanceados. Alternativamente, o limiar poderia ser ajustado para maximizar diretamente o benefício líquido, aumentando ainda mais o Recall, ou para priorizar eficiência relativa (ROI) em contextos de capacidade de investigação limitada. Essa flexibilidade de calibração é uma vantagem importante do sistema, permitindo adaptação a diferentes prioridades estratégicas e contextos operacionais.

A utilização de modelos baseados em árvores (*ensemble*), embora mais complexos que modelos lineares, foi viabilizada pela integração com técnicas de XAI (SHAP), oferecendo simultaneamente alta acurácia preditiva e transparência para auditoria — aspecto crítico em ambientes regulados como o mercado segurador.

4.5.3 Implicações para a prática

Os resultados sugerem diretrizes concretas para implementação. Primeiramente, sinistros em que o segurado é culpado (`Fault_Policy_Holder = 1`) devem ser priorizados para investigação, especialmente quando combinados com alta taxa histórica de fraude da apólice (`BasePolicy_fraud_rate`), ausência de terceiro culpado e padrões temporais atípicos. A transparência proporcionada pelo SHAP permite que analistas compreendam os fatores de risco específicos de cada caso, facilitando a tomada de decisões informadas.

O sistema permite calibração dinâmica do limiar de decisão conforme disponibilidade de recursos: limiares mais altos reduzem o volume de investigações (maior Precision), enquanto limiares mais baixos maximizam a captura de fraudes (maior Recall). Essa flexibilidade operacional possibilita adaptação a diferentes contextos de negócio. Recomenda-se monitoramento contínuo das métricas em produção e retreinamento periódico, dada a evolução natural dos padrões de fraude ao longo do tempo (*concept drift*).

O sistema complementa, mas não substitui, processos de investigação existentes. As predições devem ser interpretadas como sinais de alerta que direcionam a atenção de analistas humanos, não como decisões automatizadas de negação de cobertura. Limitações metodológicas, incluindo a utilização de parâmetros de custo médios e a ausência de valores

exatos de indenização no *dataset*, serão discutidas em maior detalhe no Capítulo 5. Em síntese, os resultados demonstram que a integração de modelos avançados de aprendizado de máquina, tratamento adequado do desbalanceamento e técnicas de interpretabilidade resulta em um sistema viável para detecção de fraudes em seguros automotivos.

5 CONCLUSÕES

Este trabalho propôs e avaliou um sistema de detecção de fraudes em seguros automotivos baseado em aprendizado de máquina, integrando desempenho preditivo, viabilidade econômica e interpretabilidade das decisões. O modelo campeão (`CatBoost + SMOTEENN`) demonstrou resultados consistentes sob múltiplas perspectivas de avaliação, evidenciando a viabilidade técnica e financeira de implementação em ambiente de produção. Este capítulo sintetiza as principais contribuições do trabalho, reconhece suas limitações e propõe direções para pesquisas futuras.

5.1 Principais Contribuições

As principais contribuições deste trabalho podem ser organizadas em três eixos: (i) contribuições acadêmicas, (ii) contribuições metodológicas e (iii) contribuições práticas para o setor segurador.

Contribuições acadêmicas:

- **Validação do Kappa de Cohen como métrica de eficiência operacional:** Demonstração empírica de que, em cenários de custos assimétricos, a otimização de hiperparâmetros guiada pelo Kappa de Cohen resulta em maior eficiência operacional (ROI) quando comparada a estratégias focadas puramente em sensibilidade. Esse achado reforça a relevância do Kappa como indicador robusto para decisões de negócio em problemas desbalanceados, complementando a literatura recente (Huayanay; Bazán; Russo, 2024).

- **Integração multidimensional de perspectivas de avaliação:** Unificação de três dimensões de análise em um único estudo — rigor estatístico (MCC, G-Mean, Kappa), viabilidade econômica (benefício líquido, ROI) e transparência decisória (SHAP) — superando a visão limitada da acurácia em dados desbalanceados e fornecendo uma avaliação holística da aplicabilidade de modelos preditivos no setor segurador.

Contribuições metodológicas:

- **Pipeline robusto com prevenção de vazamento de dados (*data leakage*):** Implementação de uma arquitetura que integra técnicas avançadas de engenharia de atributos — incluindo *target encoding*, detecção de anomalias via *Isolation Forest* e taxas de fraude agregadas por categoria — garantindo que estatísticas globais sejam

calculadas estritamente no conjunto de treino. Essa abordagem serve como referência técnica para projetos de aprendizado de máquina em ambientes de produção.

- **Score composto para seleção robusta de modelos:** Desenvolvimento de uma métrica agregada (MCC, G-Mean, Kappa) para seleção do modelo campeão, evitando a escolha de algoritmos que maximizam uma única dimensão de desempenho às custas de outras igualmente relevantes. Essa estratégia equilibra captura de fraudes, especificidade e confiabilidade das predições.

- **Otimização Bayesiana com ajuste integrado de limiar de decisão:** Utilização de *Optuna* com *threshold tuning* incorporado à função-objetivo, reforçando que o ajuste do limiar de decisão é tão crítico quanto a seleção de hiperparâmetros algorítmicos para o desempenho final do sistema.

- **Resolução de incompatibilidades críticas em ecossistemas modernos de aprendizado de máquina:** Identificação e documentação de conflitos de versionamento entre bibliotecas amplamente utilizadas no estado da arte (`xgboost`, `shap` e `numpy` 2.0) em ambiente Windows, os quais impediam a execução conjunta do treinamento e da interpretabilidade do modelo. O trabalho propôs e validou uma arquitetura baseada em ambientes virtuais segregados (`tcc_env` para modelagem e `shap_env` para explicabilidade), contornando erros de desserialização e inconsistências de *feature names*. Essa solução fornece um guia prático e reprodutível para pesquisadores e profissionais que utilizam combinações modernas dessas bibliotecas em projetos de detecção de fraude.

Contribuições práticas:

- **Demonstração quantitativa de viabilidade econômica:** Evidência empírica de que a modelagem preditiva estruturada gera valor econômico mensurável em operações antifraude. O estudo demonstrou que estratégias aleatórias (classificador *dummy*) geram ROI marginal de 156%, enquanto o modelo proposto atinge ROI de 943% e benefício líquido de R$ 3.508.000, representando ganho de 23,9% sobre o *baseline* linear e multiplicando o benefício absoluto em aproximadamente 11 vezes em relação à estratégia aleatória.

- **Transparência decisória via SHAP:** Abertura da estrutura interna do modelo baseado em árvores, revelando que a culpa do segurado (`Fault_Policy_Holder`) é o preditor individual mais importante, com importância SHAP substancialmente superior à segunda variável mais relevante. Essa descoberta oferece insumos acionáveis para equipes de investigação e contribui para a conformidade com requisitos de governança e *compliance* no mercado segurador.

5.2 Limitações

O sistema desenvolvido apresenta limitações que devem ser consideradas em implementações práticas e que oferecem oportunidades para extensões futuras. Primeiramente, os parâmetros de custo adotados (R$ 40.000 para fraude não detectada e R$ 1.000 para investigação) são valores médios representativos, mas não refletem a variabilidade real dos custos de sinistros e investigações. Em implementações operacionais, recomenda-se calibrar esses parâmetros com dados históricos da seguradora e conduzir análises de sensibilidade abrangentes para avaliar a robustez das conclusões econômicas sob diferentes cenários de custo.

Adicionalmente, o *dataset* utilizado (*Fraud Oracle Dataset*) não contém valores exatos de indenização, limitando a precisão das estimativas econômicas. A simulação de custos médios, embora válida para fins de demonstração metodológica, não captura a heterogeneidade real dos sinistros. Por fim, embora o modelo tenha sido avaliado rigorosamente em conjunto de teste independente e via validação cruzada estratificada, a avaliação em dados de outras seguradoras, regiões geográficas ou contextos regulatórios distintos seria necessária para confirmar a generalização dos resultados. Essas limitações não comprometem a validade das conclusões no escopo deste estudo, mas destacam a importância de validações adicionais em contextos operacionais específicos.

5.3 Trabalhos Futuros

Os resultados deste trabalho abrem caminho para extensões promissoras em duas direções principais, organizadas de forma sequencial.

Linha 1: Validação com dados reais de seguradoras brasileiras. A primeira prioridade para trabalhos futuros consiste na aplicação do *pipeline* desenvolvido a dados proprietários de seguradoras brasileiras, contemplando bases de grande porte com valores reais de indenização. Sugere-se a condução de estudos em colaboração com grandes companhias do setor para avaliar a generalização dos resultados em um contexto regulatório, cultural e operacional distinto do observado no *dataset* de referência (de origem internacional). Essa validação permitiria investigar se os padrões identificados, em especial a importância dominante da variável de culpa do segurado (`Fault_Policy_Holder`), se mantêm no mercado local, bem como refinar as estimativas econômicas com base em custos operacionais reais.

Linha 2: Análise de sensibilidade econômica e sistemas adaptativos. Com base nos dados reais obtidos na Linha 1, propõe-se a realização de uma análise de sensibilidade sistemática variando a razão entre o custo da fraude (C_F) e o custo de investigação (C_I). O objetivo seria mapear a relação entre a estrutura de custos do portfólio e a métrica de otimização mais adequada, investigando a existência de um ponto de inflexão

a partir do qual a maximização do benefício líquido absoluto (favorecida por métricas como o MCC) supera a necessidade de eficiência percentual (favorecida pelo Kappa). Do ponto de vista prático, essa análise poderia subsidiar o desenvolvimento de **sistemas de decisão adaptativos** em seguradoras, nos quais a política de investigação seria calibrada dinamicamente conforme o valor do sinistro: sinistros de alto valor (razão C_F/C_I elevada) adotariam limiares mais agressivos priorizando captura de fraudes, enquanto sinistros de varejo (razão C_F/C_I baixa) adotariam limiares conservadores priorizando eficiência operacional. Essa abordagem conectaria definitivamente a modelagem de aprendizado de máquina à estratégia financeira da companhia, transcendendo a busca isolada por acurácia preditiva.

Essas duas linhas de pesquisa, organizadas sequencialmente, representam uma evolução natural e estruturada do trabalho realizado, com potencial de impacto significativo tanto acadêmico quanto prático no setor segurador.

REFERÊNCIAS

ABDALLAH, A.; MAAROF, M. A.; ZAINAL, A. Fraud detection system: A survey. **Journal of Network and Computer Applications**, Elsevier, v. 68, p. 90–113, 2016.

ABI. **Insurance Fraud Statistics 2022**. [*S.l.*], 2022. Disponível em: https://www.abi.org.uk/products-and-issues/topics-and-issues/fraud/.

AKIBA, T. *et al.* Optuna: A next-generation hyperparameter optimization framework. *In*: **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**. [*S.l.: s.n.*], 2019. p. 2623–2631.

ALTALHAN, F. A.; ALGARNI, A. Imbalanced data problem in machine learning: A review. **IEEE Access**, v. 13, p. 12535–12551, 2025.

ARRIETA, A. B. *et al.* Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. **Information Fusion**, Elsevier, v. 58, p. 82–115, 2020.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations**, v. 6, n. 1, p. 20–29, 2004.

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. **Statistical science**, Institute of Mathematical Statistics, v. 17, n. 3, p. 235–255, 2002.

CAIF. **The Impact of Insurance Fraud on the U.S. Economy**. [*S.l.*], 2022. Disponível em: https://insurancefraud.org/wp-content/uploads/The-Impact-of-Insurance-Fraud-on-the-U.S.-Economy-Report-2022-8.26.2022.pdf.

CHAWLA, N. V. *et al.* Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, Springer, v. 21, n. 1, p. 6, 2020.

CNSEG. **22º Ciclo do Sistema de Quantificação de Fraudes (SQF) – Relatório Completo 2024**. [*S.l.*], 2024. Disponível em: https://cnseg.org.br/publicacoes/22-ciclo-do-sqf-completo-2024.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

DERRIG, R. A. Insurance fraud. **Journal of Risk and Insurance**, Wiley Online Library, v. 69, n. 3, p. 271–287, 2002.

DING, N. *et al.* Automobile insurance fraud detection based on pso-xgboost model and interpretable machine learning method. **Insurance: Mathematics and Economics**, Elsevier, v. 120, p. 51–60, 2025.

HE, H. *et al.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In*: **Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN)**. [*S.l.: s.n.*], 2008. p. 1322–1328.

HUAYANAY, A. d. l. C.; BAZÁN, J. L.; RUSSO, C. M. Performance of evaluation metrics for classification in imbalanced data. **Computational Statistics**, Springer, v. 40, n. 3, p. 1447–1473, 2024.

ITRI, B. *et al.* Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *In*: IEEE. **2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)**. [*S.l.: s.n.*], 2019. p. 1–6.

KOMSRIMORAKOT, P. *et al.* Enhancing fraud detection in imbalanced motor insurance claims using a novel hybrid resampling technique. **Journal of Big Data**, Springer, v. 12, n. 1, p. 1–25, 2025.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in artificial intelligence**, Springer, v. 5, n. 4, p. 221–232, 2016.

KUBAT, M. Addressing the curse of imbalanced training sets: one-sided selection. *In*: MORGAN KAUFMANN. **Proceedings of the 14th international conference on machine learning**. [*S.l.: s.n.*], 1997. p. 179–186.

LUNDBERG, S. M. *et al.* From local explanations to global understanding with explainable ai for trees. *In*: **Nature Machine Intelligence**. [*S.l.: s.n.*], 2020. v. 2, p. 56–67.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. **Biochimica et Biophysica Acta (BBA)-Protein Structure**, Elsevier, v. 405, n. 2, p. 442–451, 1975.

SAARELA, M.; PODGORELEC, V. Recent applications of explainable ai (xai): A systematic literature review. **Applied Sciences**, MDPI, v. 14, n. 19, p. 8884, 2024.

SHAPLEY, L. S. A value for n-person games. **Contributions to the Theory of Games**, v. 2, n. 28, p. 307–317, 1953.

.1 Lista Completa de Atributos do Dataset

Este apêndice apresenta a lista completa dos 33 atributos do *Fraud Oracle Dataset*, incluindo a variável-alvo. Os atributos estão organizados em ordem alfabética e classificados quanto ao tipo de dado (categórico ou numérico). A Tabela 5 descreve cada atributo, seu tipo, valores possíveis e significado no contexto de sinistros de seguros automotivos.

Tabela 5 – Descrição completa dos atributos do *Fraud Oracle Dataset*

Atributo	**Tipo**	**Descrição**
`AccidentArea`	Categórico	Área onde ocorreu o acidente. Valores: *Urban* (urbana), *Rural* (rural).
`AddressChange_Claim`	Categórico	Tempo desde a última mudança de endereço do segurado. Valores: *no change*, *under 6 months*, *1 year*, *2 to 3 years*, *4 to 8 years*.
`Age`	Numérico	Idade do segurado em anos. Intervalo: 0 a 80 anos.
`AgeOfPolicyHolder`	Categórico	Faixa etária do segurado. Valores: *16 to 17*, *18 to 20*, *21 to 25*, *26 to 30*, *31 to 35*, *36 to 40*, *41 to 50*, *51 to 65*, *over 65*.
`AgeOfVehicle`	Categórico	Idade do veículo em anos. Valores: *new*, *2 years*, *3 years*, *4 years*, *5 years*, *6 years*, *7 years*, *more than 7*.
`AgentType`	Categórico	Tipo de agente que vendeu a apólice. Valores: *External* (externo), *Internal* (interno).
`BasePolicy`	Categórico	Tipo base da apólice. Valores: *Liability* (responsabilidade civil), *Collision* (colisão), *All Perils* (todos os riscos).
`DayOfWeek`	Categórico	Dia da semana em que ocorreu o acidente. Valores: *Monday*, *Tuesday*, *Wednesday*, *Thursday*, *Friday*, *Saturday*, *Sunday*.
`DayOfWeekClaimed`	Categórico	Dia da semana em que a reclamação foi registrada. Valores: *Monday*, *Tuesday*, *Wednesday*, *Thursday*, *Friday*, *Saturday*, *Sunday*, *0* (valor especial).
`Days_Policy_Accident`	Categórico	Intervalo de dias entre a emissão da apólice e o acidente. Valores: *none*, *1 to 7*, *8 to 15*, *15 to 30*, *more than 30*.
`Days_Policy_Claim`	Categórico	Intervalo de dias entre a emissão da apólice e a reclamação. Valores: *none*, *8 to 15*, *15 to 30*, *more than 30*.

Continua na próxima página

Tabela 5 – *Continuação da página anterior*

Atributo	**Tipo**	**Descrição**
`Deductible`	Numérico	Valor da franquia da apólice em dólares. Valores possíveis: 300, 400, 500, 700.
`DriverRating`	Numérico	Classificação do motorista pela seguradora. Escala: 1 (melhor) a 4 (pior).
`Fault`	Categórico	Parte responsável pelo acidente. Valores: *Policy Holder* (segurado), *Third Party* (terceiro).
`FraudFound_P`	Numérico	**Variável-alvo.** Indica se o sinistro é fraudulento. Valores: 0 (legítimo), 1 (fraude).
`Make`	Categórico	Marca do veículo. 19 valores possíveis, incluindo *Honda*, *Toyota*, *Ford*, *Mazda*, *Chevrolet*, *Pontiac*, *Accura*, *Dodge*, *Mercury*, *Jaguar*, *Lexus*, *Porche*, *Saab*, *Saturn*, *BMW*, *Nisson*, *VW*, *Ferrari*, *Mercedes*.
`MaritalStatus`	Categórico	Estado civil do segurado. Valores: *Single* (solteiro), *Married* (casado), *Widow* (viúvo), *Divorced* (divorciado).
`Month`	Categórico	Mês em que ocorreu o acidente. Valores: *Jan*, *Feb*, *Mar*, *Apr*, *May*, *Jun*, *Jul*, *Aug*, *Sep*, *Oct*, *Nov*, *Dec*.
`MonthClaimed`	Categórico	Mês em que a reclamação foi registrada. 13 valores possíveis (12 meses + valor especial *0*).
`NumberOfCars`	Categórico	Número de veículos segurados pelo segurado. Valores: *1 vehicle*, *2 vehicles*, *3 to 4*, *5 to 8*, *more than 8*.
`NumberOfSuppliments`	Categórico	Número de suplementos ao sinistro. Valores: *none*, *1 to 2*, *3 to 5*, *more than 5*.
`PastNumberOfClaims`	Categórico	Número de reclamações anteriores do segurado. Valores: *none*, *1*, *2 to 4*, *more than 4*.
`PoliceReportFiled`	Categórico	Indica se foi registrado boletim de ocorrência policial. Valores: *Yes* (sim), *No* (não).
`PolicyNumber`	Numérico	Número identificador único da apólice. Intervalo: 1 a 15.420.
`PolicyType`	Categórico	Tipo completo da apólice (combinação de categoria do veículo e cobertura). 9 valores possíveis, incluindo *Sport - Liability*, *Sport - Collision*, *Sport - All Perils*, *Sedan - Liability*, *Sedan - Collision*, *Sedan - All Perils*, *Utility - Liability*, *Utility - Collision*, *Utility - All Perils*.

Continua na próxima página

Tabela 5 – *Continuação da página anterior*

Atributo	**Tipo**	**Descrição**
`RepNumber`	Numérico	Número identificador do representante da seguradora. Intervalo: 1 a 16.
`Sex`	Categórico	Sexo do segurado. Valores: *Male* (masculino), *Female* (feminino).
`VehicleCategory`	Categórico	Categoria do veículo. Valores: *Sport* (esportivo), *Sedan* (sedan), *Utility* (utilitário).
`VehiclePrice`	Categórico	Faixa de preço do veículo em dólares. Valores: *less than 20000*, *20000 to 29000*, *30000 to 39000*, *40000 to 59000*, *60000 to 69000*, *more than 69000*.
`WeekOfMonth`	Numérico	Semana do mês em que ocorreu o acidente. Valores: 1, 2, 3, 4, 5.
`WeekOfMonthClaimed`	Numérico	Semana do mês em que a reclamação foi registrada. Valores: 1, 2, 3, 4, 5.
`WitnessPresent`	Categórico	Indica se havia testemunhas no acidente. Valores: *Yes* (sim), *No* (não).
`Year`	Numérico	Ano em que ocorreu o acidente. Valores: 1994, 1995, 1996.

Observações:

- O dataset contém 24 atributos categóricos e 8 atributos numéricos (excluindo a variável-alvo `FraudFound_P`).
- Alguns atributos categóricos representam variáveis originalmente numéricas que foram discretizadas em faixas (por exemplo, `AgeOfVehicle`, `VehiclePrice`, `Days_Policy_Accident`).
- O atributo `PolicyNumber` é um identificador único e não possui valor preditivo direto para detecção de fraudes.
- O atributo `RepNumber` identifica o representante da seguradora responsável pela venda da apólice.
- A variável-alvo `FraudFound_P` é binária, onde 0 indica sinistro legítimo e 1 indica fraude.

A CÓDIGO-FONTE

A.1 Repositório GitHub

Todo o código-fonte, dados, modelos treinados e documentação técnica deste trabalho estão disponíveis publicamente em um repositório GitHub, garantindo total reprodutibilidade dos resultados apresentados.

A.1.1 Acesso ao Repositório

URL: https://github.com/edurodrigues-usp/tcc-fraud-detection-autoinsurance

Figura 9 – QR Code para acesso direto ao repositório GitHub.

Nota: O QR Code acima pode ser escaneado com qualquer aplicativo de câmera de smartphone para acesso direto ao repositório.